

A loss of vision

Appropriations bills drawn up this summer suggest that Republicans in the House of Representatives have been paying little more than lip-service to the importance of a balanced science budget.

Anyone who thought that Washington had got the message about the importance of science funding, and had reached a bipartisan consensus on long-term federal investment in research, will have received a rude awakening in these few past weeks, as the US House of Representatives has considered the appropriations bills that will fund government spending in the next fiscal year.

The House appropriations bills, which are being drawn up under tight 'spending caps' negotiated back in 1997, are far from the last word in determining these funding levels for 2000. But they still constitute the most explicit statement available of the priorities of this chamber of the Republican Congress.

Science and technology account for one dollar in seven of the so-called discretionary government spending that springs from the 13 appropriations bills. When Newt Gingrich was Speaker of the House, he instructed the appropriations subcommittees that basic science in particular was to be protected, even when money was tight. With Gingrich gone, however, some of these subcommittees appear to see science as a soft touch for reductions in funding.

At the Department of Energy, although House appropriators have supported non-weapons physics programmes, the proposed Spallation Neutron Source would get only a quarter of the money needed to begin construction. Plans for a new scientific simulation initiative at the energy department civilian laboratories would be shelved. (Across the whole government, in fact, the House cuts would eliminate most of the \$360 million information-technology research programme proposed in February by the Clinton administration).

Over at the commerce department, the House is seeking to eliminate the Advanced Technology Programme at the National Institute of Standards and Technology (NIST). Nor is there any money either for construction of a long-planned Advanced Measurement Laboratory at NIST.

The worst news for science, however, comes in the appropriations bill for Veterans' Affairs, Housing and Urban Development (VA/HUD), which, for complex historical reasons, includes funding for the National Science Foundation (NSF) and the space agency NASA. The NSF gets no increase at all from the VA/HUD bill passed by

the appropriations committee (see *Nature* 400, 490; 1999). NASA is hit for \$1 billion in this bill, with space science cut 8 per cent from last year and Earth science programmes 16 per cent. According to a survey carried out by *The Chronicle of Higher Education*, each of these bills has also been larded with an unprecedented number of district-specific, pork-barrel research projects aimed at securing the bill's passage in a chamber where the Republicans hold the narrowest of margins.

The research agency missing from this status report is, of course, the National Institutes of Health (NIH). That's because NIH is in the appropriations bill for labour, health and education, to be drawn up by a subcommittee chaired by John Porter (Republican, Illinois). Porter's bill has been held back until last, while its total allocation has been subjected to recurrent raids by the other 12 appropriations bills that went before it, and now stands 20 per cent short of what its programmes cost this year.

That leaves the Congress facing an almighty budget crunch when it returns in September. It can either repeal the budget caps explicitly, fiddle the figures to find Porter's shortfall or implement massive cuts on politically popular programmes. Some predict that the caps will be bypassed before the new fiscal year begins on 1 October, and surging tax revenues are duly redirected to finance everyone's favourite programmes. But such a benign outcome looks increasingly elusive. The caps are, after all, the Republican party's main political achievement after four-and-a-half years in control of the Congress. And the demands for extra money have become so overwhelming that they are, overall, insatiable.

With the support of Porter and others, biomedical research may well find itself among the lucky few programmes whose demands are met when the dust settles. Thanks largely to choices by House appropriators, other science programmes will be less fortunate. That doesn't point to the balanced science policy recommended last year in a report from Vernon Ehlers (Republican, Michigan), deputy chairman of the house science committee, to his colleagues, as best serving the interests of the United States. Republicans who wish to restore the balance could start by amending the VA/HUD bill when the full House considers it after the summer recess. □

Summer-time blues

French researchers, angry and upset over last week's synchrotron decision, deserve a full explanation.

Seasoned French politicians are well aware that the best time for unpopular announcements is August, when virtually the whole country shuts down and goes on holiday. Claude Allègre, the embattled minister for national education and research, appears to be no exception. At least not to judge by the timing of last week's controversial announcement that France is to become a full partner in the construction and operation of a new 3-GeV synchrotron facility, Diamond, being planned in Britain (see page 604).

There is much to commend this decision, even though it has angered a broad swathe of researchers — enthusiastically backed by local political leaders — who had been hoping to see their own proposed synchrotron, Soleil, built on French soil. Allègre is right to argue

that economies of scale alone make it sensible that large research facilities should increasingly be planned at a European level. They should not be seen as public-works projects to mop up local unemployment or to burnish the electoral prospects of local politicians.

But the critics, too, have their points. Collaborating with Britain will save on the short-term costs of building two separate machines; but longer-term savings are more questionable. And many French researchers face the substantial costs of travelling hundreds of miles to carry out experiments that others will be able to do on their doorstep. Allègre owes his critics a full and detailed description of the calculations behind his decision. This is unlikely to placate them; but it may show that the decision has more logic than many now accept. □



Researcher sues university over rights to transgenic technology

[SAN DIEGO] A legal dispute has blown up over the rights to a potentially valuable technique developed at the University of Hawaii for introducing foreign genes into living organisms.

Anthony Perry, a molecular embryologist formerly at the Babraham Institute in Britain who has been working with Ryuzo Yanagimachi's renowned mouse-cloning team at Hawaii, is suing the university over exploitation of the novel transgenic technique. The technology could have broad applications in genetically engineering mammals to produce human substances (see *Nature* **394**, 408; 1998).

In a lawsuit filed in state court in Honolulu last month, Perry alleges that the university improperly licensed the technology to a small biotechnology firm without his permission. He has asked the court to rule on the university's actions and to rectify alleged violations.

Walter Kirimitsu, Hawaii's general counsel, says that the university holds exclusive patent rights and title to the transgenic technology, and licensed it appropriately. The university will respond to Perry's lawsuit in court next week, and a judge will consider the legal arguments at subsequent hearings.

Although disputes involving patent rights and universities are not uncommon, the case could stir considerable controversy, given the volatility of public attitudes towards genetic engineering and cloning experiments.

The disputed technology is called 'mamalian transgenesis by intracytoplasmic sperm injection', or ICSI transgenesis. It was reported in a paper in May in *Science* by the so-called 'Team Yana', with Perry as the lead author (see *Science* **284**, 1180–1183; 1999).

Last year, the team reported full development of mice from a cloning technique (see *Nature* **394**, 369–374; 1998). Although the current legal dispute centres on the novel transgenic method, it may in future envelop the team's cloning process as well.

In court records, Perry claims that it was he who invented the transgenic technology, along with other novel methods employed by the team. The technology can be used to genetically engineer animals — such as pigs, cattle or sheep — to produce human hormones or proteins, or drugs.

Typically, a US university owns the rights to any technology developed by its scientists, while the patent bears the name of the scientific inventor once the lengthy filing process is completed. The university can negotiate with a company or organization for develop-

ment of a technology. The university and scientific inventors then share the royalties from any commercial product.

In the current dispute, however, Perry — who was not employed by the university at the time — says in court records that the university licensed the transgenesis technology without his consent. He also questions the university's title to the technology, and seeks title to the technology for development through a company with which he is associated.

The circumstances have been complicated by the fact that Perry's work in Hawaii was carried out while he was a European Molecular Biology Organization fellow. Another key author of the *Science* and *Nature* papers is Teruhiko Wakayama, who was in Hawaii as a fellow of the Japanese Society for the Promotion of Science before joining the university late last year.

After leaving his post at the Babraham Institute in Cambridge, Perry sought to join the university as well. But his hiring has been suspended pending the outcome of the dispute. Funding for Yanagimachi's laboratory comes from the US National Institutes of Health.

The University of Hawaii licensed tech-

nological inventions from Yanagimachi's lab to ProBio, a company headed by Australian businessman Laith Reynolds of Perth. Ironically, it was Perry who first proposed a university deal with Reynolds. But Reynolds and Perry have since fallen out, reportedly after Perry became upset that the role of the scientists involved in the early experiments was being minimized.

Reynolds calls Perry's legal action "disingenuous", and adds: "We are continuing with our commercial activities while the university sorts out its problems with one of its staff."

Perry and Wakayama have since formed their own company, BiogeneSys International LLC, with funding from John Henry Felix, a member of the board of trustees of the Salk Institute in La Jolla, California. Perry and Felix — a well-connected Honolulu politician and former owner of a chain of mortuaries — have sought to add Salk scientists to their firm's scientific advisory board.

As the civil lawsuit began, Perry's attorney asked a Hawaii administrative court to declare void the university's deal with ProBio, arguing that it violated state procurement law. An administrative judge is expected to rule soon on that request. **Rex Dalton**

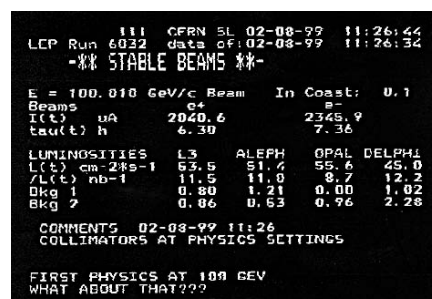
CERN collider homes in on Higgs boson

[MUNICH] Scientists at the European Laboratory for Particle Physics (CERN) are celebrating a breakthrough which some claim could in principle allow them to produce and begin to characterize the elusive Higgs boson before the end of this year.

Over the past three years, CERN's Large Electron-Positron Collider (LEP) in Geneva, the world's most powerful particle accelerator, has been tuned to deliver higher and higher energy beams. Last week, almost exactly ten years after its first experiments, LEP was pushed beyond its design limits when its colliding beams of electrons and positrons achieved energies of 200 GeV (individual beam energies of 96 GeV were reached in May, and 98 GeV in July).

This achievement could, say some physicists, be sufficient to detect the Higgs boson, the particle physicist's Holy Grail. Recent measurements at LEP predict a mass of around 109 GeV for the standard model Higgs boson — a particle whose existence was hypothesized to explain why all other particles have such a wide range of masses.

Not everyone is confident that the Higgs mass will be quite so low, and the boson may therefore remain elusive until the



Into the record book: LEP report on the first collisions at 200 GeV centre-of-mass energy.

construction of CERN's Large Hadron Collider. But the response among CERN scientists when LEP hit its energy target "was quite noisy", says Patrick Janot, the collider's physics coordinator. "We opened all our champagne bottles at the same time."

Mike Witherell, new director of the US particle physics laboratory Fermilab, describes it as a "great achievement to be able to squeeze so much physics out of LEP". He adds that the Higgs mass is "so fundamental to our field" that every increase in GeV is very important to home in more precisely on the likely mass. **Alison Abbott**

Coordinated work reveals plant evolution

[ST LOUIS] A 'tree of life' giving the most complete picture yet of the evolutionary relationships among all of the Earth's green plants was unveiled last week at the International Botanical Congress in St Louis, Missouri.

The scenario was drawn up by the Green Plant Phylogeny Research Coordination Group, a five-year project involving around 200 scientists from a dozen nations. It is also known as 'Deep Green'.

The team found that plants are divided into four separate kingdoms: green plants — the largest group with about 500,000 species, including all land plants and some aquatic plants such as green algae — brown and red seawater plants, and fungi. Interestingly, the fungus kingdom is the closest relative of the animal kingdom, life's fifth kingdom.

Previously, scientists divided life on Earth into animal and plant kingdoms. But the results of the Deep Green project show that, for plants, there are "four lineages of complex, nucleated organisms".

Deep Green researchers also found that all green plants appear to come from a single lineage, not from multiple lineages as previously believed. "This indicates that there's an Eve — a common ancestor — in the primordial soup of green plants," says Brent D. Mishler, a professor of integrative biology at the University of California, Berkeley, and a co-principal investigator of the team.



Trees of life: 'plants exist as four kingdoms'.

"A better understanding of this tree of life will allow scientists to better predict the biological properties of plants," says Mishler. "Their economic importance as sources of medicines, structural materials, food and chemicals is immense."

The Deep Green project was described as a classic example of the benefits of interdisciplinary cooperation between scientists, from taxonomists to molecular biologists. "Instead of people jealously guarding their data, they shared data and produced many important discoveries," says Mishler.

Systematists produced data on the evolutionary origin of individual plants, for instance, then molecular biologists refined the species' position on a cladistic tree. This effort "makes a science out of the field of taxonomy", adds Mishler. "Since Darwin, people have been speculating on phylogenetics. This effort gives us methods and explicit data to make comparisons for green plants."

The results of Deep Green's work were presented at eight symposia during the week-long Botanical Congress, which is held every six years. More than 4,000 scientists attended.

One unexpected discovery was that plant life on land derived from fresh, not salt, water. "This overturns the traditional thinking among scientists and what is taught in every textbook in America," says Mishler.

As universities and schools begin to teach Deep Green scientific findings, controversy may arise in the United States, since the discoveries are based on the principles of natural evolution. Indeed, some of those at the meeting warned that creationists may try to use the textbook revisions to advance their cause.

Deep Green was funded by the US National Science Foundation, the Department of Energy and the Department of Agriculture. The findings are available on: <http://ucjeps.berkeley.edu/bryolab/greenplantpage.html>. Congress reports are available at: www.abc99.org. **Rex Dalton**

CORBIS/JIM ZUCKERMAN

'Strengthened' Icelandic bioethics committee comes under fire

[MUNICH] The government of Iceland has been accused of contravening an international convention by setting up a national bioethics committee whose members are selected exclusively by government departments.

The health minister announced last week that the new committee will immediately replace Iceland's existing seven-member bioethics committee. The latter was set up in 1997, and its members were selected by the ministry from nominations put forward by the academic and medical communities.

Members of the new five-member committee will be selected without external nominations. Its main duties will be to approve research protocols involving patients — in particular, patient data held in a planned national health database which will be run by a private company — and to give general advice to politicians on the ethical aspects of medical treatments.

Officials defend the new arrangement for selecting committee members by saying that the new members would be free to speak their minds. But critics say it could mean that only individuals who support official policies are appointed.

Pétur Hauksson, vice-president of Mannvernd, the Association of Icelanders for Ethics in Medicine and Science, argues that the move contravenes the 1964 Helsinki agreement on Medical Research on Humans, which states that all research ethics committees should be independent of the researchers, those who initiate or finance the research and the relevant authorities. In this case, the government is the relevant authority because it is in charge of licensing out the database.

Hauksson is particularly concerned that a committee chosen by government officials "may not be able to give truly objective views and opinions on new government plans to further regulate health-sector databases and biobanks, and to revise personal privacy laws".

He also claims that the change was prompted by the government's desire to remove from the committee critics of the private company deCODE Genetics. The company is expected to be granted an exclusive 12-year licence by the government to create and market a national database combining genetic, health-care and genealogical information about Iceland's

small population (see *Nature* 396, 395; 1998).

Ingibjörg Pálmadóttir, Iceland's Minister of Health, and Sigurdur Gudmundsson, the director-general of public health, reject such criticisms. Pálmadóttir says the judgement of the bioethics committee will not be biased just because its members are selected by the government, and that the new regulations will strengthen the committee's power.

Vilhjálmur Arnason, director of the University of Iceland's Institute of Ethics, said on public radio last weekend that the change in selection procedures would not strengthen the committee, but would "strengthen the control of the government over the committee". He claimed that it was "almost too obvious" that the grounds for the change were the protection of the financial interests in the database project.

But Bogi Andersen, an Icelandic molecular geneticist working at the University of California, San Diego, and a vociferous critic of deCODE's plans for running the national database, says that parts of the new regulations are very positive, such as the likelihood that the bioethics commission would include a lay representative. **Alison Abbott**

UK space strategy sets sights on Mars

[LONDON] Britain's science minister, Lord David Sainsbury, last week announced a shift in the government's space science policy, to give more money to basic research at the expense of large Earth observation programmes.

The decision paves the way for a UK-led Mars lander project, designed to search for evidence of life on the planet.

In presenting the new strategy at the Science Museum in London, Sainsbury announced a redistribution of £19.5 million (US\$31.3 million) of the £180 million UK budget for space expenditure. About half of this will be used for the European Space Agency's (ESA) Artes programme, a government-industry partnership for advanced telecommunications technologies.

The rest of the money will support basic space research and the development of basic technology, which many scientists complain has been neglected.

Britain's space scientists welcomed the government's agreement to contribute £5 million to the estimated £25 million development cost of a Mars lander spacecraft, Beagle 2. The ESA has selected Beagle 2 as part of the payload of its 2003 Mars Express mission, provided that the research team can find sufficient financial support.

The Particle Physics and Astronomy Research Council (PPARC) has agreed to provide £2.8 million for the instruments for Beagle. The research team, led by Colin Pillinger, director of the Planetary Sciences Research Institute at the Open University in Milton Keynes, has also raised £12 million from private sponsorship.

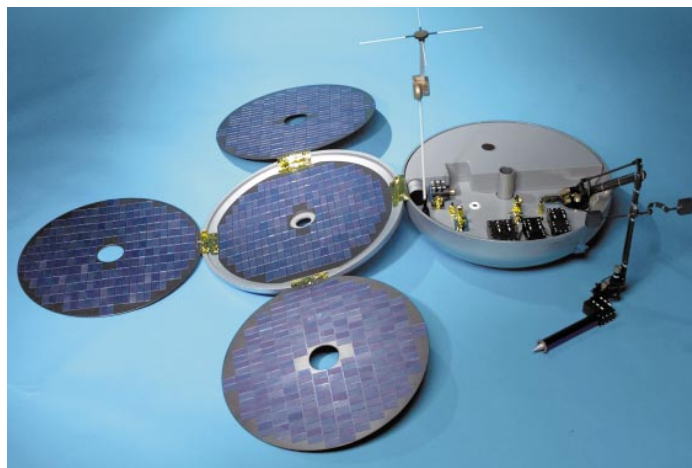
Beagle 2 will carry out in-situ analyses of the Martian atmosphere and geology. Unlike the US space agency NASA's Pathfinder, Beagle will have devices to retrieve material from beneath the planet's surface, where organisms could most likely survive.

It will also be capable of measuring the isotopic composition of excavated rocks, rather than just their chemistry. This is essential for understanding the nature of the processes that created the material.

The project's supporters argue that, even if no evidence of life is discovered, the mission could provide a valuable scientific service in settling the controversial question of whether life ever existed on Mars.

"If biological processes have left traces in Martian material, than Beagle is bound to find evidence," says Paul Merdin, director of science at the British National Space Centre.

Not all planetary scientists share this view. Some call the goal of the mission a "long shot" that would rely on luck. To answer conclusively questions about life on Mars would require a major drilling rig capable of excavating material from 30 to 300



Beagle bound for Mars: the lander will excavate the planet's surface in search of evidence that life could have existed on the 'red planet'.

BEAGLE2.OPEN.AC.UK

metres depth, they argue, rather than a robot capable only of "scraping the surface".

But both the ESA and its industrial partner, Matra Marconi Space, have confirmed the potential value of the lander following a review of the technical and scientific aspects in April. "Beagle 2 would offer unprecedented scientific opportunities," says Roger Bonnet, ESA space science director.

Pillinger's team is still short of several million pounds. This means a race against time. Beagle's financing must be secure by next January if it is to be approved as part of the Mars Express mission.

Alan Wells, director of the space research centre at the University of Leicester, which is involved in the project, predicts that Sainsbury's announcement will "serve as a catalyst for bringing in other sources of funding". Pillinger is reported to have "very good prospects" of obtaining additional funds.

The new space strategy also backs Britain's contribution to the ESA's infrared cornerstone mission First and to the microwave background cosmological mis-

sion Planck, scheduled to launch together in 2007. PPARC has allocated £22 million to fund instruments for the two missions.

Sainsbury highlighted Britain's major involvement in ESA's Living Planet programme for environmental sciences, a planned series of small and cheap Earth observation missions. Britain will lead the first mission, Cryosat, which will monitor the impact of global warming on the polar ice caps.

Underlining a trend towards smaller and more focused missions, the UK contribution to large-scale ESA Earth observation projects has been cut by almost 25 per cent, to £44 million. But this has been compensated for by an increase in national programmes.

UK scientists have generally welcomed the plans. Wells agrees that the strategy is "a move in the right direction". But he criticizes the fact that the overall public space budget is only being redirected rather than expanded. "Private investment in space technology exceeds public expenditure by a factor of three," he says. "But no returns to basic research come back from industry."

Quirin Schiermeier

Legal threat in bid to ban xenotransplants

[WASHINGTON] A US advocacy group that opposes xenotransplantation threatened last week to sue the Department of Health and Human Services (DHHS) for failing to respond to a legal petition demanding a ban on the practice because of its potential health risks.

The New York-based Campaign for Responsible Transplantation (CRT) filed a petition in December asking the DHHS to ban animal-to-human transplants of cells, tissues and organs.

The legal deadline for the department's response was 10 June. If the DHHS does not respond by 17 August, says CRT, the group will sue for breach of the Administrative Procedures Act. This law allows courts to compel government agencies to take action

when this has been "unreasonably delayed".

Lorrie McHugh, a DHHS spokeswoman, says, "We want to be as responsive as possible to the campaign". She says that the DHHS is preparing an interagency response to the group's "very complicated legal document". The respondents include the Food and Drug Administration and the National Institutes of Health.

CRT says that xenotransplantation poses unacceptable risks because of the danger of the transfer of deadly animal viruses to humans. In April, the Food and Drug Administration announced a *de facto* ban on xenotransplants from non-human primates, but left the way open for the development of transplants from pigs (see *Nature* 398, 549; 1999).

Meredith Wadman

US acts to lengthen term of patents to help biotechnology

[WASHINGTON] In a boost to the US biotechnology industry, the House of Representatives last week approved a bill that would reverse the reduction in patent protection that companies say they are facing as a result of 1994 legislation that implemented the General Agreement on Trade and Tariffs (GATT).

The bill, sponsored by Congressman Howard Coble, (Republican, North Carolina), passed by 376 votes to 43. It aims to mitigate the effects of a GATT provision that changed patent terms to 20 years from the date a patent application is filed. Patent protection was previously granted for 17 years from the date a patent was issued.

The biotechnology industry says the 1994 change shortened the terms of its patents, as biotechnology inventions usually take longer to win approval at the Patent and Trademark Office (PTO) than other inventions, which take less than two years on average.

Under the bill — the American Inventors Protection Act — inventors would be compensated by having a day added to the patent term for each day over three years that it takes the PTO to grant a patent (providing that the delay is not the applicant's fault). This would effectively guarantee a 17-year patent term from the date a patent is issued.

Chuck Ludlam, the vice-president for government relations at the Biotechnology Industry Organization, says it is "an incredibly important provision" to restore patent terms "that otherwise would be eroded".

The bill also requires the PTO to publish after 18 months any patent applications that are already published in other countries. Industry had lobbied for publication of all applications, not just those published abroad. It says that publication helps companies to avoid pursuing research and development in areas already staked out by others.

But independent inventors are opposed to this provision, arguing that companies might intimidate them or steal their inventions if they are published before patents are granted.

Although the bill has not been introduced in the Senate, Jeanne Lopatto, a spokeswoman for the judiciary committee, says that the committee's chairman, Senator Orrin Hatch (Republican, Utah), "is committed to patent reform" and will address the issue in Congress.

The bill does not address the issue of 'first to file' versus 'first to invent', which divides the countries of the European Commission from the United States (see *Nature* 397, 457; 1999). Some European countries were seeking a US shift on this in exchange for accepting a period of grace on patent applications in Europe.

Meredith Wadman

French anger over plan to back British synchrotron

[PARIS] Last week's decision by the French government to support the construction of a new synchrotron facility in Britain rather than France (see *Nature* 400, 489; 1999) has generated a strong backlash in the French scientific community.

The day after the announcement, Robert Comes, director of the Laboratoire de l'Utilisation du Rayonnement Électromagnétique (LURE) at Orsay, south of Paris, and its two deputy directors, Jean Daillant and Abderrahmane Tadjeddine, issued a statement saying they would refuse to cooperate with the plan.

Before last week's decision there had been hopes that the French government would replace the ageing facilities at LURE with Soleil, a 2.15-GeV third-generation synchrotron (see *Nature* 392, 114; 1998).

One of the main complaints is that Diamond, the planned British 3-GeV synchrotron, will fall short of the needs of the 2,000 French scientists who rely on such facilities for their experiments. "From a scientific point of view, the common project will at best fill one-quarter to one-third of French needs, and probably about one-half of British needs," the LURE directors say.

Their complaints are echoed by members of a joint committee between the Atomic Energy Commission and the Centre National de Recherche Scientifique which was set up to consider synchrotron radiation in France. Also, SNCS, the main trade union representing scientists, and scientists at Orsay have expressed anger at the decision made by Claude Allègre, the minister of education, research and technology.

Other opponents of Allègre's decision include politicians from six regions that were eager to sponsor half the cost of building a synchrotron facility in their district. They are seeing thousands of potential jobs disappear, together with the economic boost and prestige of such a large-scale facility.

French scientists have pushed for Soleil for more than a decade. But Allègre announced on 2 August that, instead of going ahead with it, the government would give an estimated FF350 million (US\$56 million) over seven years towards Diamond, and FF60 to FF80 million a year to its operating costs.

"From its arrival, the French government has made known that it wished all future large scientific facilities to be carried out on a European scale," Allègre stated.

Ironically, only the Greens of the Ile-de-France welcomed the decision. "The European partnership and the pooling of facilities, finances and skills represent the future," says Michel Michelon, president of the



Rough Diamond? French scientists say new UK synchrotron (above) will not meet their needs.

national and regional development committee for the Ile-de-France council.

Allègre's opponents disagree with the cost analysis said to be behind his decision. "It's not at all obvious that the English solution is less expensive," says Yves Farge, a former ministerial adviser and former head of research at Pechiney, who wrote a detailed report on Soleil.

Farge and other scientists say that Allègre based his decision on a report drawn up by one of his advisers, Paul Clavin, who is not a known expert in the field. "What needs to be done is to make the report public," Farge says.

Clavin admits that he is not an expert on synchrotrons, but argues that the costs of Soleil would be "excessive for France", adding that its initial designs did not focus sufficiently on the needs of structural biology. "Also, it would be irresponsible to let the chance of collaborating with the British pass by."

Yves Petroff, director-general of the European Synchrotron Radiation Facility at Grenoble, who was asked to review Clavin's document by the minister, says: "Basically, everything is incorrect. I have no objection to cooperating with the British or any other country, but, based on this report, that is totally irrational."

Petroff argues that Allègre underestimates the cost of participating in the British synchrotron, and says that France will eventually end up spending the same money that it would have done in building Soleil.

Scientists and politicians say they are also upset that Allègre announced the decision at the beginning of August, when many French scientists are on vacation. Allègre is expected to explain his decision in more detail at a meeting scheduled for September.

Jean-Paul Huchon, president of the Ile-de-France region, had called for the meeting with Allègre. Staff at LURE and other groups will then assess their course of action. Until then, says Chantal Damais of the SNCS's national bureau, there will be little activity. "The scientists are extremely indignant — but they are at the beach." **Heather McCabe**

Japan plans to label 'detectable' GM food

[TOKYO] Japan's Ministry of Agriculture, Forestry and Fisheries last week unveiled draft regulations stipulating the mandatory labelling of 28 food products containing detectable genetically modified (GM) ingredients, including soya beans, maize and potatoes.

The rules were approved by the ministry's committee on GM food labelling on Tuesday (10 August). Any foods with 5 per cent or more DNA or protein from GM crops would have to be labelled.

The regulations also specify that products resulting from the mixture of GM and non-GM foodstuffs should be labelled as 'undifferentiated'. Products that lose protein and genetic material from GM crops in the production process, such as soy sauce, would not need labelling (see *Nature* 395, 628; 1998).

The ministry's plan, which will come into effect next April, has provoked strong criticism from industry and consumer groups. Japanese companies are concerned that the regulations will lead to trade friction with exporters of GM foods, in particular the United States. They also fear that the prices of their products will be driven up because of

the high cost of carrying out tests to identify foods containing GM elements.

Many food manufacturers plan to switch to suppliers that provide GM-free crops, but this will not be easy. Japan imports almost 60 per cent of its agricultural products, mainly from the United States, where identification of GM ingredients is not required. Switching to domestically produced GM-free ingredients could double the cost of products, according to industry analysts.

"You have to raise the price of the products, and one wonders whether this is what consumers want in exchange for having a few GM foods labelled," says Yoshinori Komura, a member of the ministry's committee on GM organisms and director of the Japan Vegetable Oil Association. He points out that there are no approved tests for detecting GM molecules in Japan.

Consumer groups, on the other hand, say the draft regulations should be more stringent if the main aim is to respect the right of the consumer to know whether a product contains GM components.

"We are not interested whether the product contains a detectable amount of GM

products or not. We want to know whether it contains such products at all," says Satoko Tanaka, a representative of civil rights groups in Tokyo.

"Labelling products with undifferentiated ingredients would only confuse the public," says Yasue Ito, director of the Consumer Science Federation. "This is just avoiding the need to track down the genetic history of the ingredients."

Ito points out that differentiation of GM and non-GM products should be made possible by the exporters isolating crops more effectively.

Japanese companies involved in the development of GM crops have become increasingly pessimistic about commercializing their products. Japan Tobacco and Sun-tory are carrying out farm-scale trials of GM crops, including virus-resistant rice. They say they have no immediate plans to release any of their products on the market.

Some companies have commercialized GM flowers, including blue carnations. But none is exploiting the food market, out of concern that this could give the company a negative image.

Asako Saegusa

Tension stays high as US optical societies launch merger ballot

[SAN DIEGO] The proposed merger of the two leading US optical-science societies has turned into a bruising political campaign as voting began last week among the more than 26,500 combined members.

As the Optical Society of America (OSA) and the International Society of Optical Engineers (SPIE) started sending out ballot papers, charges of unfair and unethical tactics were already flying.

The infighting is concentrated within OSA, whose 12,500 members are largely academic scientists. Proponents of the merger, who include the leaders of both organizations, argue that it will strengthen the role of optics science.

But critics fear that scientists will get short shrift if merged with SPIE's 14,000 members, who are primarily engineers. Some OSA scientists see the merger "as an attack on their livelihood" as it could diminish their role in the society, for which they earn credit that advances their academic careers.

They are accusing the organization's leaders of trying to improperly influence the election. As an example of inappropriate leadership tactics, they cite the fact that OSA's executive director recently contacted SPIE members offering OSA membership at half price along with the opportunity "to cast an historic vote" in the OSA election.

Among some rank and file OSA members



there is deep resentment towards the leadership's stance. "It has become really reprehensible," says Kenneth Alexander, a visual psychophysicist at the University of Illinois at Chicago's College of Medicine.

As chairman of the OSA's division of vision and colour, Alexander recently tried to e-mail a brief notification of the merger vote to the approximately 400 scientists in his division. But the e-mail was blocked by Anthony E. Siegman, president of OSA and an emeritus professor of electrical engineering at Stanford University.

Another OSA scientist was threatened with legal action by the society's executive director for using a society mailing list to send out a notice on the merger.

Siegman rejects allegations that he and other advocates of the merger are trying to influence the election result. "The initiative for unification is the product of a very long, careful and arduous effort by distinguished OSA members" to present all sides of the issue, he says, pointing out that repeated OSA board votes have overwhelmingly approved the merger proposal.

OSA's membership solicitation to SPIE members was a routine recruitment effort, Siegman says, denying that it was aimed at manipulating voting. He adds that limiting the use of the society's mailing list is in line with a long-standing policy to protect society members from unnecessary correspondence.

Paul Foreman, chairman of OSA's merger task force and a leading advocate of the merger, accuses some of the critics of paranoia, claiming that they have disseminated false information about the merger and are impossible to satisfy. What they are doing is "a disservice to the profession", he says.

There appears to be little ferment within SPIE, although its treasurer, Charles DeMund, voted against the merger "because it is like two people, who are not getting along while dating, getting married".

The OSA election will be held on 29 September during its annual meeting in Santa Clara, California.

Rex Dalton

Protests at Indian solar observatory refit

[NEW DELHI] The Indian Institute of Astrophysics (IIAP) in Bangalore has triggered a fierce controversy with a plan to convert part of the 100-year-old Kodaikanal Solar Observatory in southern India into a modern auditorium that will host an international scientific meeting.

The meeting, supported by the International Astronomical Union, will be held in December as part of the centenary celebrations for the observatory, which sits on the Palani Hills. Its six-inch telescope has been monitoring the Sun since 1899.

It was at Kodaikanal, 90 years ago, that John Evershed discovered the phenomenon of radial motion in sunspots — the Evershed effect has only recently been explained (see *Nature* **389**, 47–49; 1997).

Around 50 foreign scientists have been invited to the meeting, one of several events planned by IIAP to mark the centenary. But some institute scientists are concerned about the potential impact of the renovation work on the historic observatory.

The roof of the main hall, the heart of the observatory connecting the two domes (pic-



Kodaikanal observatory: preparing to host an international meeting to mark its centenary.

tured above), has already been dismantled to make way for the auditorium. According to one scientist, a stone pier that supported the man-sized Shelton astronomical clock has been thrown away and the clock has been put in another building.

The Kodaikanal observatory was set up in British-ruled India by meteorologists, not physicists. According to Rajesh Kochhar, an IIAP astronomer and a science historian, this followed “a famine in the Madras presidency which underscored the need for a study of the Sun so that monsoon

patterns could be better understood”.

According to Kochhar, throughout its 100 years of existence, the Kodaikanal observatory has been an important international research facility, cooperating with observatories in Paris, Greenwich, Mount Wilson and Cambridge.

IIAP director Ramnath Cowsik denies any “damage to the building or the instruments”. He says that the renovation work is being carried out in accordance with the suggestions of the Archaeological Survey of India, and that any displaced instruments have been preserved in another laboratory along with the Shelton clock.

IIAP is fully funded and owned by the central government through the Department of Science and Technology. Valangimman Ramamurthi, the department’s secretary, says that damage to the observatory is being exaggerated. He adds that the department is prepared to have the renovation work examined by a team from the Indian National Trust for Arts and Cultural Heritage, the body responsible for the preservation of historic sites.

K. S. Jayaraman

Animal rights activists attack Gore over chemical screening

[WASHINGTON] Animal rights activists launched a series of television advertisements last week attacking US Vice-President Al Gore for backing a controversial government programme to screen common chemicals for toxicity.

People for the Ethical Treatment of Animals (PETA) is running the advertisements in New Hampshire and Iowa, which are key political states for Gore, a presidential hopeful. In the advertisements, television actress Bea Arthur says that Gore wants to poison and kill millions of animals “just to prove things like turpentine and rat poison are dangerous”.

The six-year, \$500–700 million programme was launched by Gore in 1998 and is being run by the Environmental Protection Agency (EPA). It asks companies to screen ‘voluntarily’ 2,800 chemicals that are made or imported in quantities of more than a million pounds a year.

The list includes a wide array of substances, from caffeine to carbon tetrachloride. So far, 218 companies have signed up to screen 1,153 chemicals. In December, the EPA will begin requiring the participation of companies that have not volunteered. The data generated will be posted publicly on the Internet.

Seventeen animal rights groups, including the Humane Society, have written to Gore arguing that many of the chemicals “clearly need no further testing”.

Other sceptics included the late congressman George Brown (Democrat, California). He wrote to Carol Browner, the EPA administrator, last December, citing concerns about “the welfare of millions of animals to be used in the testing” and that ample data already exist for many of the chemicals.

The critics argue that Gore and the EPA, by using high production volume as a surrogate marker for human and environmental exposure, are needlessly targeting chemicals that are either benign, or toxic but already well characterized.

Procter & Gamble, which is participating in the programme, wrote to Browner in January saying: “We are concerned that EPA will be perceived as ‘setting back the clock’ by failing to apply thoughtful scientific approaches” to reducing the use of animals.

Katherine Stitzel, associate director of the company’s human safety department, argues that the agency is “just looking at whether the material is toxic at the largest dose possible to an animal. We don’t think that is germane to exposure to the environment and humans.”

Others in industry support the programme’s use of animals, including the Chemical Manufacturers Association. The association’s spokesman Tom Gilroy says: “Companies don’t want to do any more [animal] testing than is necessary. It is expensive. But they want their data to be of

sufficient quality that it’s accepted.”

Gore, environmentalists and the EPA argue that there is a dearth of publicly available information about the chemicals, and that some animal sacrifice is justified in the pursuit of public health and environmental protection.

The EPA says that full screening data, defined by the Organization for Economic Cooperation and Development, are available for only seven per cent of the 2,800 chemicals.

“The bottom line goal is to protect children and adults from being exposed to chemicals that may be harmful,” says Chris Lehane, Gore’s press secretary.

William Sanders, director of the EPA’s Office of Pollution Prevention and Toxics, estimates that “a fraction of a million” animals, mostly rats, will be killed under the programme. The EPA initially estimated that up to two million animals would die. But it later revised its recommendations on how companies should test, proposing the use of some non-animal alternatives, and that tests on chemicals of similar structure should be consolidated.

Alan Goldberg, director of the Johns Hopkins Center for Alternatives to Animal Testing, says the EPA has been “remarkably responsive in trying to follow a humane approach”. But animal activists argue that the agency cannot require companies to minimize the use of animals because the programme is voluntary.

Meredith Wadman

Endocrine disrupter verdict left open

[WASHINGTON] Insufficient data exist to provide an accurate assessment of the risks to public health posed by the chemicals known as endocrine disrupters in the low ambient doses found in the environment, according to the US National Research Council (NRC).

A panel of experts put together by the council, part of the National Academies complex, has agreed that some of the chemicals can be shown to adversely affect reproduction and development of wildlife at high doses.

But, after a drawn out and highly-charged four-year investigation, the members of the panel disagreed about the extent to which the concentrations of the chemicals found in the environment constitute a public health risk.

The release of the study — which leaves the endocrine disrupter question unresolved — was broadly welcomed by both sides in the debate, each being left free to interpret the outcome as they saw fit.

Theo Colburn of the Worldwide Fund for Nature (WWF) argues that “this is a clear message that there’s enough evidence that we should be taking precautions, and acting to fill gaps in our knowledge”. Colburn’s book, *Our Stolen Future*, expounded the endocrine disrupter hypothesis.

But the Chemical Manufacturers Association said the study “found little evidence to support certain allegations about chemicals and adverse effects in people”. It said the report was “a thorough work that provides solid direction for continued research”.

Proponents of the endocrine disrupter hypothesis, such as the WWF, argue that some chemicals mimic oestrogens and other sex hormones in ways that can have extremely adverse effects, even in doses far too small to trigger a conventional toxic response.

Some proponents attribute many types of cancer, as well as an alleged global decline in human sperm counts, to the effect. Because of disagreements on the NRC panel about the validity of the hypothesis, it chose to call the chemicals ‘hormonally-active agents’.

The panel recommends more research to test the endocrine disrupter hypothesis, including long-term studies of exposed populations. But it concedes that even this may fail to bridge the bitter divisions between adherents to the hypothesis and sceptics. “Some of the differences reflect areas where additional research would help, others reflect differing judgements about the significance of the existing information,” says the summary of the report, released on 3 August.

Ernst Knobil, an endocrinologist at the University of Texas at Houston and chair of the panel, warns that these differences are likely to persist. Controls will be difficult to find for long-term human studies because the chemicals are so pervasive in the environment, he says. And the banning of production of some



At risk? Eggs of the least tern — an endangered species and possible victim of PCBs — being checked in Gulfport, Mississippi.

of the most active chemicals, such as polychlorinated biphenyls (PCBs), will gradually reduce their concentration in the environment, making their effects even more elusive.

Policy-makers say that the report could lend impetus to research programmes. The report “may help to get the research programme back on track,” says a White House

aide who follows the issue. The Clinton administration has proposed that the Environmental Protection Agency should spend \$21 million on research into the subject next year, up from \$16 million this year. A House of Representatives appropriations subcommittee has proposed adding a further \$5 million.

But the report is less than enthusiastic about the quality of existing research. “In much of the literature, distinctions between environmental concentration, exposure and dose are often not made,” it says. “This failure can reduce the relevance and value of research results.”

The panel agreed that “many wildlife studies show associations between reproductive and developmental anomalies and exposure to environmental contaminants, some of which are hormonally-active agents”. It also said that there is evidence that prenatal exposure to high concentrations of some of the chemicals “can affect the developing nervous system”.

It found that the literature did not support associations between adult exposure to the chemicals and breast cancer or other hormonally sensitive cancers. But it noted that few studies had checked concentrations of these chemicals in the body against cancer risk in adults; none had checked for the effect of exposure to hormonally-active agents during fetal life on cancer incidence later on.

The study also found that, because of wide and unexplained geographical variations, the current data do not support (or refute) the widely-propagated theory that overall human sperm counts have been falling.

Colin Macilwain

Brain drain accelerates from Siberia

[MOSCOW] The exodus of trained scientists from the Siberian branch of the Russian Academy of Sciences is accelerating. A new report claims that last year the academy lost 668 scientists, including 384 with doctorates; these figures are up from 535 and 288 in 1997.

The report, prepared by the local Centre for Social Adaptation and the Philosophy and Law Institute of the Siberian branch of the academy, points out that only five to ten per cent of those scientists who left the academy went abroad, but that those who did are among the best in their fields.

The survey lists several reasons for the exodus. The first, named by almost 95 per cent of respondents, is the higher salaries that scientists receive abroad. The second, given by 80 per cent, is the declining prestige of science in Russia.

The third reason for leaving is a desire to get closer to new ideas, conferences and other

scientific events. The final reason, identified by one third of respondents, is a fear for the future of children in Russia.

According to the report, Russia has paid a high cost for the emigration of its scientists, who were expensive to train. But it points out that those who leave Russia but stay in science are not lost completely as they often maintain links with former colleagues. Most say they would like to return when the situation in Russian science improves.

More damaging is the fact that many talented scientists move into business and other work unrelated to science.

The final loss to Russian science, says the report, is the low productivity of scientists experiencing harsh living conditions and low morale. One in four scientists at the Siberian branch of the academy live below the official poverty line, and over 40 per cent have a below-average standard of living. **Carl Levitin**

Congress backs plan for agency to run nuclear weapons labs

[WASHINGTON] Representatives of both chambers of the US Congress have agreed to a major reorganization of the Department of Energy that would create a National Nuclear Security Administration within the department to run its nuclear weapons laboratories.

The reorganization — Congress's main response to allegations of spying by China at the Los Alamos National Laboratory in New Mexico — is designed to give the administrator of the new agency direct authority to manage the weapons laboratories.

Bill Richardson, the energy secretary, believes that the reorganization will restrict his control over the weapons laboratories. He may ask President Bill Clinton to veto it if, as expected, both houses of Congress pass it in September as part of a military spending authorization bill. But the president may be reluctant to intervene because the bill's passage is needed to assure the operation of the Pentagon.

Richardson, meanwhile, has appointed Thomas Gioconda, an air-force general, to replace Victor Reis as acting assistant

energy secretary in charge of the weapons laboratories.

Australia and NZ bring in labelling for GM foods

[SYDNEY] Australia and New Zealand are to follow the European Union and, more recently, Japan in requiring the labelling of food containing genetically modified products (see page 605). The health ministers of Australia and New Zealand announced last week that they have "agreed to require mandatory labelling of foods produced using gene technology and those containing genetically modified ingredients".

Final decisions on aspects such as methods of labelling and the threshold level for detecting protein and GM ingredients will be made by the Australia–New Zealand Food Standards Council in October.

Digging deep for French nuclear waste lab

[PARIS] The French government last week gave permission for an underground laboratory for the study of nuclear waste disposal to be built in Bure in Meuse, northeastern France. It will cost FF1.5 billion (US\$245 million) over five years. Construction work will start in the autumn,

with research scheduled to begin in 2001.

Local community groups and environmentalists are challenging the move, saying the government is preparing a nuclear waste storage facility. Other critics want the government to reconsider its decision, and favour storage at the surface or subsurface near nuclear centres. A 1991 law states that parliament must agree on a method of waste storage in 2006.

Spain climbs towards the science top ten

[BARCELONA] Spain's scientific output is growing faster than anywhere else in the world, both in terms of quantity of publications and their quality, according to a report released by the Higher Council of Scientific Research.

Spain increased its output of publications from 0.97 to 2.37 per cent of the global total between 1981–85 and 1993–97, climbing from sixteenth to eleventh place internationally. Overall scientific output increased by 3.5 times during this period, compared to a global average of 1.4 times.

Spain managed to increase the impact factor of its output — a measure of how widely publications are cited — by a factor of 2.12 over this period, followed by Austria (1.75), Ireland (1.63) and Finland (1.54). But

the report points out that the impact factor of Spanish publications is still 21 per cent less than the global average, which rose by a factor of 1.31.

The Spanish research area with the highest international output is astrophysics. Fields with an impact factor equal to or higher than the average for the European Union include anthropology and computing sciences.

Argentina and Brussels agree research pact

[LONDON] Research data from Argentina and member states of the European Union are to be exchanged under an agreement on scientific and technological cooperation struck between Brussels and Buenos Aires. The deal allows research teams from Argentina to take part in the first phase of the fifth Framework programme of research, which was agreed in principle last year.

Cooperation will cover the fields of the quality of life and management of living resources, the information society, competitive and sustainable growth, and energy, environment and sustainable development. It would be carried out by the pooling of research projects and exchanges of information, visits and exchanges of scientists, joint organization of conferences, and the sharing of equipment and materials.

Robots make it a World Cup double for France

[PARIS] Perhaps inspired by their country's real-life victory in the soccer World Cup last summer, French computer engineers walked away as champions at this year's Robot World Cup Soccer Games. In the final held last week in Stockholm, the French team from the Laboratoire de Robotique de Paris beat Australia's University of New South Wales 4–1 in the canine-like quadruped league.

The category features robots, programmed to score goals, that measure 30 centimetres long and 20 centimetres wide.

From 2002, RoboCup will feature a bipedal, humanoid robot league. And researchers predict that, by 2050, a team of humanoid robots will have the skills to beat the human soccer world champions.

Clinton's technology programme faces axe

[WASHINGTON] The House of Representatives has passed an appropriations bill that would eliminate all funding for the Advanced Technology Programme (ATP) at the National Institute of Standards and Technology (NIST). President Bill Clinton had proposed spending \$240 million next year on ATP, a technology support

programme which he strongly favours.

Republicans in Congress have been trying to close down the programme for several years, but had backed off after reforms were implemented in a bid to satisfy critics. This year's House appropriations bill for the Department of Commerce — the latest of several bills to cut science and technology spending — would eliminate ATP.

Were appetites eclipsed down on the farm?

[LONDON] This week's total eclipse in the Northern Hemisphere may have helped to further the frontiers of scientific research in at least one unexpected direction — by revealing what controls the grazing patterns of dairy cattle. Britain's Institute of Grassland and Environmental Research planned to use the eclipse to test the theory that daylight determines when cattle eat.

The research will take place at a farm in Cornwall, where the eclipse is due to last for two minutes and two seconds. An average cow feeds three to four times a day, with its longest meal lasting three to four hours and finishing around sunset, regardless of time of year. The eclipse takes place when light levels are normally near their maximum, so, if the theory is correct, it should have a large effect on cows' eating patterns.

Why quasi-persons are not patentable

Sir — You report that the US patent office has rejected a bid by Stuart Newman and Jeremy Rifkin to patent techniques for creating a human–animal hybrid (*Nature* 399, 626; 1999). In rejecting their claim, the examiner informed the would-be inventors that the US Patent Act does not cover the patenting of humans. Although the examiner is correct that property rights in a human cannot be granted under US law, the Newman–Rifkin invention raises a different question, as only a part of it is human. But even something that is only part human — a quasi-person — cannot be patented, according to constitutional law.

To be granted a patent in the United States, an invention must be new, useful and not obvious to a person skilled in the pertinent scientific discipline. In 1976, Chakrabarty tried to patent a bacterium genetically engineered to digest oil but, although the bug satisfied all three criteria, the patent office refused to grant a patent on the grounds that living organisms could not be patented. The Supreme Court overturned this decision, concluding that “Anything

man-made under the sun can be patented”.

That opened the door for patents on engineered, living organisms. Did it make humans patentable? A human being is ‘man-made’ (woman-made, too, of course) so it satisfies the first level of inquiry set out by the Supreme Court. A human is also new (the combination of genes is novel), useful, and not obvious (no one could predict what maternal and paternal genes would come together nor the resulting phenotype). All the criteria to earn a patent are met.

But the matter is not settled so simply. Patented inventions become the personal property of the inventors to dispose of as they see fit. However, under US law, a human cannot be held as property. The thirteenth amendment to the constitution, which ended slavery, prohibits taking property rights in human beings. A human, therefore, cannot be patented.

What about the patentability of hybrids? Rifkin suggests that the answer requires the patent office to define the maximum quantity of humanness that can be present in an organism for it to be considered

patentable. But this analysis has already been done. Not only does the thirteenth amendment prohibit the ownership of a person, but it also forbids property rights in a quasi-person. When the constitution was framed in 1787, it contained a provision that slaves were to be counted as three-fifths of a person for the purpose of apportioning representatives to the Congress. The goal of the amendment was to outlaw slavery — to bar the ownership of Black people. But it did not change the legal status of Black people. They remained part-people (quasi-people) who were not entitled to vote, citizenship, or other rights. It was not until the fourteenth amendment was passed that Black people were made citizens and granted full personhood under the constitution.

The thirteenth amendment stood for the proposition that a quasi-person was not property to be held by another. So the patent office is correct in refusing to grant a patent on an organism that is only part human.

Richard M. Lebovitz

*Evogen, 3515 Washington Blvd, No. 317,
Arlington, Virginia 22201, USA*

Where a little aid could go a long way

Sir — I support all the matters raised in the Commentary by Hin and Subramaniam, “Scientific societies build better nations” (*Nature* 399, 633; 1999). The suggestion that aid agencies should provide money to support scientific societies in developing countries, and to create federations of societies, strikes a chord. Even the production of a newsletter in Africa costs money that is not easily raised.

I am concerned with the development of biochemistry in Eastern Europe and Africa, and participated in the launch of the Federation of African Societies of Biochemistry and Molecular Biology. Many of the problems are similar in these different regions. The authors say that the creation of societies depends on the “coming together of scientists committed to translating a vision into reality”. This is true but too often such people are inhibited because there is a bad relationship between the government and the universities. In these circumstances it takes a brave person to found a society that may later ask the government to increase funds for science. Such people tend to be regarded as troublemakers.

Effective aid should involve a partnership between the aid agency and the government but too often they act

independently. In the absence of an indication that the government takes science seriously, aid to individuals will often merely encourage a brain drain.

Peter Campbell

University College London, London WC1E 6BT, UK

More evidence needed on effect of ivory ban

Sir — Payne *et al.* express concern at the resumption of a limited trade in elephant ivory from Namibia, Zimbabwe and Botswana to Japan, and at the system for monitoring the illegal killing of elephants (MIKE) that supports the decision by the Convention on International Trade in Endangered Species (CITES) to allow this trade (*Nature* 399, 297; 1999). The authors correctly identify difficulties in devising survey methods, and of relating changes in populations to underlying causes. They are also disturbed that this decision may lead to proposals for future sales. But they misrepresent or misunderstand key points.

CITES did not decide to resume a limited ivory trade. Before the 1989 ban there had been a rampant ivory trade from illegally killed elephants. CITES decided to allow a one-off limited trade this year in stockpiled ivory under tightly controlled conditions, and to monitor the impact of this trade as an

‘experiment’. Now that the sales have taken place, no further legal trade is allowed.

The 1989 ban should also have been regarded as an experiment, and monitored. When CITES meetings in 1992 and 1994 questioned whether the ivory ban was working effectively, no monitoring system was in place to provide an informed response. MIKE aims to fill a major gap by providing the best possible objective information for decision-makers.

The MIKE system was designed by IUCN, the World Conservation Union, as required by CITES. Experts were brought together from Africa and Asia, and their proposed system was passed through IUCN’s elephant specialist groups for refinement, before approval by the CITES standing committee. What further peer review would Payne *et al.* desire?

An objective assessment of the one-off sale is not yet available. Nor has it been possible to assess the impact of the ban in any meaningful quantitative way. Yet Payne *et al.* assume that a total ban is the safest option for elephants, despite the numbers killed illegally since 1989. MIKE has been designed to inform debate on such issues. If there is insufficient understanding to make decisions about limited trade, the same applies to decisions about a continued ban.

Nigel Leader-Williams

*Durrell Institute of Conservation and Ecology,
Department of Anthropology, University of Kent at
Canterbury, Canterbury CT2 7NS, UK*

The second Silent Spring?

The drive to squeeze ever more food from the land has sent Europe's farmland wildlife into a precipitous decline. How can agricultural policy be reformed so that we have fewer grain mountains and more skylarks?

**John R. Krebs, Jeremy D. Wilson,
Richard B. Bradbury and Gavin M.
Siriwardena**

The Roman poet Catullus lamented the death of a pet sparrow (*"Passer mortuus est meae puellae"*), probably a house sparrow (*Passer domesticus*) or, just conceivably, a tree sparrow (*P. montanus*). In either case, a present-day Catullus would have ample cause for lamentation. Both kinds of *Passer* are among the many species of birds, invertebrates and plants that have declined dramatically in Northern Europe during the past 25 years.

We estimate that, in the past 20 years, ten million breeding individuals of ten species of farmland birds have disappeared from the British countryside. For example, the corn bunting (*Emberiza calandra*) and tree sparrow have declined for periods of up to a decade at an average rate of more than 5% per year. The declines in bird numbers in part reflect those in the invertebrate and plant populations upon which they depend.

Parallel changes have taken place in many other European countries, although these have not been documented in as much detail as in Britain, where censuses are carried out by the British Trust for Ornithology (BTO). In all, 116 species of farmland birds — one fifth of European avifauna — are now of conservation concern. Such is the significance attached to the loss of bird populations that John Prescott, UK Deputy Prime Minister and environment secretary, has included a composite trend of 139 species in the list of 13 headline indicators of sustainable development upon which the government reports annually.

Where have all the birds gone?

Rachel Carson's classic 1963 book *Silent Spring* alerted the public to the toxic side effects of the organochlorine insecticides, such as DDT, that had fuelled the green revolution. Residues of these pesticides were found to persist in the food chain, reaching higher concentrations, and hence having more severe effects, at successive trophic levels. Famously, they were identified as the cause of rapid population decline in birds of prey, such as the peregrine falcon and sparrowhawk, through the thinning of eggshells. The offending chemicals have now been phased out in the United Kingdom and many other countries, but their use is still increasing in some parts of the world.

The new losses in biodiversity are sometimes called the 'second Silent Spring'. However, although they are associated with the



Menaced by monoculture: the numbers of British farmland birds, such as (from left to right) corn bunting, skylark, house sparrow, lapwing, linnet and yellowhammer, are falling rapidly.

intensification and industrialization of agriculture¹, they involve more subtle and indirect effects than the poisoning of wildlife by pesticide residues. In general terms, intensification is about making as great a proportion of primary production as possible available for human consumption. To the extent that this is achieved, the rest of nature is bound to suffer.

Detailed ecological studies have shown the devastating effect of the intensification of agriculture on biodiversity. Here we summarize some of the key results, taking birds as our illustrative taxon, and link them to the broader issues of society's choices about the kind of landscape and environment it wishes to bequeath to future generations.

Can we be sure that the bird declines in the United Kingdom are caused by agricultural intensification? Most of the evidence is by association, but in sum total it is damning. For example, annual BTO censuses of 42 species of breeding birds show that 13 species living exclusively in farmland, such as the skylark (*Alauda arvensis*) and corn bunting, declined

by an average of 30% between 1968 and 1995, while 29 species of habitat generalists, such as the carrion crow (*Corvus corone*) and the wren (*Troglodytes troglodytes*), have increased by an average of 23% (ref. 2). More direct evidence is that the declines of four species have actually been reversed, at least on a local scale, by 'experimental' changes in farming: by the grey partridge (*Perdix perdix*; by reduced pesticide input, provision of nesting cover and predator control)³, ciril bunting (*E. cirilus*; provision of overwinter stubbles and grass margins)⁴, corncrake (*Crex crex*; provision of early nesting cover and delayed harvest of grass for forage)⁵ and stone curlew (*Burhinus oedipus*; adjacent fields of spring crops and grazed pasture)⁶. Most of these manipulations would not make it into a textbook on experimental design, but they are about as strong as the evidence is likely to get that agriculture can be modified for the benefit of wildlife.

Drowning in food

The changes in British agriculture over the past 30 years, which have many parallels with other parts of the world, have sought to increase production and productivity (Table 1). The success of the green revolution in achieving this is undeniable: in spite of rapid population growth, about 25% more food per capita is produced now than 30 years ago.

In the European Union (EU), a large incentive to increase production has been the Common Agricultural Policy (CAP), which has subsidized production and kept prices artificially high. The CAP costs the EU tax-payer more than 40 billion euros

Table 1 Key changes in British agriculture over the past 30 years

Land drainage
Hedgerow removal
Introduction of new crop types (oilseed rape, linseed)
'Improvement' of pasture (fertilizers, monoculture)
Increased agrochemical inputs
Switch from spring to autumn sowing
Silage harvest of grass
Reduction in traditional rotations
Reduction in undersown leys
Taken from ref. 7.

(US\$43 billion) per year — more than half the total EU budget — and generates vast surpluses, politely called 'intervention stocks'. If the CAP continues on its present course, it will generate annual surpluses of 58 million tonnes of cereals and 1.5 million tonnes of beef by early next century. Wine and milk lakes tell a similar story. In England, subsidies vary from year to year, but from 1994 to 1998 they ranged between 45% (1994) and 123% (1998) of the 'total income from farming'. In 1998, income was £2.17 billion (US\$3.5 billion) and subsidy was £2.67 billion — an extraordinary situation in a country where state subsidy for almost all industries has disappeared.

The European Commission recognizes that this cannot go on. First, production subsidies, typically at 20%, are inimical to the principles of free trade embodied in the World Trade Organization. Second, the expansion of the EU to include the agricultural countries of Eastern Europe would break the bank and/or generate even larger surpluses if the CAP were to continue on its present course. Third, the need to conserve wildlife in harmony with agriculture is beginning to be recognized. For these reasons a reform of the CAP, called Agenda 2000, was agreed in March 1999. Agenda 2000, presumably a step on a road to much larger reforms, does not do away with subsidies to farmers, but it facilitates the switch of subsidies away from production to flat-rate payments per hectare of land (a process initiated in earlier reforms). It is argued that this will reduce the incentive for production and allow other imperatives, such as environmental benefits, to come into play. But the proposals as they stand are virtually silent about what environmental benefits are expected and how they will be achieved. What would conservation ecologists recommend?

Conserving biodiversity

The United Kingdom has three agricultural schemes that could have benefits for biodiversity. Environmentally Sensitive Areas and the Countryside Stewardship Scheme (or regional equivalents) both pay farmers to preserve traditional landscape features; between them they encompass about 12.5% of agricultural land. Unfortunately, there are few data to demonstrate whether or not these schemes have benefited biodiversity, although some habitats, such as lowland heath, have been preserved or restored. Set-aside (farmers paid under the CAP to leave land fallow to reduce surplus production) accounts for about 10% of arable farmland; the available data show that certain kinds of set-aside can be beneficial for birds and other wildlife. But set-aside will probably be phased out early in the next century, so its benefits will disappear.

Although we can, as described above, devise plans that help individual species to recover, there is no magic bullet with which

Table 2 **Beneficial effects typical of organic farming**

Taxon	Organic compared with conventional farming	Number of farms studied
Earthworms ⁸	Greater numbers and biomass	1*
Spiders ⁹	More individuals and species	3
Butterflies ¹⁰	Non-pest species more abundant	16
Arable weeds ¹¹	Greater cover and more species	8
Birds ¹²	31 out of 34 species more abundant	62
Birds ¹³	8 out of 18 species more abundant (others no difference)	44

*One site with 12 experimental plots

to reverse the declines of a large suite of species. The changes listed in Table 1, in different combinations for different species, have all brought about declines, and therefore the most general prescription is to reverse the intensification of agriculture.

A slightly more specific prescription arises from the general result that heterogeneous landscapes are beneficial for birds; for example, where crops intermingle with hedgerows or field margins, or where a diversity of crops grow in close juxtaposition. Although several comparisons of organic and conventional farms have suggested that organic farming is good for biodiversity (Table 2), this benefit probably relates to general features of habitat heterogeneity and lower-intensity agriculture, rather than to any specific prescriptions in the theology of organic farming. There have been no systematic comparisons of the biodiversity benefits of organic and other 'wildlife friendly' farming methods.

There are three unexplored research areas that may help to guide future policy for managing the agricultural landscape: the benefits of structural heterogeneity at different spatial scales (field, farm, landscape); the trade-offs between conservation benefits and the economic profitability of farming; and the trade-offs (or synergies) between conservation of different taxonomic groups. To make research results more useful for policy-makers, there is a need for integrated, predictive modelling.

On an EU-wide scale, there are unresolved questions for conservation ecology about the relative merits of moving to less intensive, more environmentally friendly agriculture throughout the countryside (the Eastern European model) versus highly intensive agriculture in bread-basket regions with separate, large nature reserves or national parks for wildlife (the North American model). The United Kingdom is probably too small for the North American model, but one could envisage some form of it on a Europe-wide basis, especially if reduced subsidies were to make agricultural production uneconomic in some areas, and instead conservation were subsidized as a 'crop'.

The next revolution?

In the United Kingdom, as in most of Europe, people have made the landscape. This means that the characteristic habitats and species that conservationists wish to preserve (such as heather moorland) are generally there

because of (traditional) land management rather than in spite of it. The future shape and purpose of the countryside is society's choice. At present, most of those in the United Kingdom who voice an opinion — from the more than one million members of the Royal Society for the Protection of Birds to John Prescott — would prefer a countryside in which agricultural production is tempered with conservation. And on a worldwide stage this makes sense for a sustainable future: the green revolution gave success at a price, and that price cannot be paid indefinitely.

The British public's concern about genetically modified (GM) crops, based in part on justifiable environmental concerns, must be placed in this context. Whatever hazard GM crops might be thought to pose to the environment is painted onto a biodiversity landscape that is already severely damaged by the intensification of agriculture. The environmental safety aspects of GM must be thoroughly investigated to define the risks before, if they prove acceptable, moving to large-scale commercial planting. Even then, there must be continued scrutiny and evaluation. But we must also recognize a potential benefit of GM crops — to give us a wider range of options as we try to make a more sustainable future for agriculture than that created by the last green revolution. □

John R. Krebs and Richard B. Bradbury are at the Edward Grey Institute of Field Ornithology, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

Gavin M. Siriwardena is at the British Trust for Ornithology, Thetford, Norfolk, UK.

Jeremy D. Wilson is at the Royal Society for the Protection of Birds, Sandy, Bedfordshire SG19 2DL, UK.

1. Pain, D. J. & Pienkowski, M. W. (eds) *Farming and Birds in Europe* (Academic, London, 1997).
2. Siriwardena, G. M. et al. *J. Appl. Ecol.* **35**, 23–43 (1998).
3. Potts, G. R. *The Partridge: Pesticides, Predation and Conservation* (Collins, London, 1986).
4. Evans, A. in *Farming and Birds in Europe* (eds Pain, D. J. & Peinkowski, M. W.) 351–357 (Academic, London, 1997).
5. Tyler, G. A., Green, R. E. & Casey, A. *Bird Study* **45**, 35–50 (1998).
6. Green, R. E. & Griffiths R. E. *J. Zool. Lond.* **223**, 457–471 (1994).
7. Fuller, R. J. in *Lowland Farmland Birds. Proceedings of the 1999 BOU Conference* (eds Aebischer, N. J., Grice, P. & Vickery, J.) (BTO, Tring, in the press).
8. Whalen, J. K. et al. *Biol. Fertil. Soils* **27**, 400–407 (1998).
9. Feber, R. E. et al. *J. Arachnol.* **26**, 190–202 (1998).
10. Feber, R. E. et al. *Agric. Ecosyst. Envir.* **64**, 133–139 (1997).
11. Moreby, S. J. et al. *Ann. Appl. Biol.* **125**, 13–27 (1994).
12. Christensen, K. D. et al. *Dansk Orn. Foren. Tids.* **90**, 21–28 (1996).
13. Chamberlain, D. E. et al. *Biol. Conserv.* **88**, 307–320 (1999).

Acknowledgements The Royal Society, NERC, BBSRC and MAFF provided support. Charles Godfray and David Gibbons commented on a draft manuscript.

From desert to deluge in the Mediterranean

Judith A. McKenzie

A time when the Mediterranean became cut off from, then reconnected to, the Atlantic has been securely dated. An approach known as astrochronology lies behind this achievement.

Some time between five and six million years ago, the Mediterranean Sea became isolated from the Atlantic Ocean. In consequence some areas dried out — hence the title of Kenneth Hsü's book *The Mediterranean was a Desert*¹ — and large salty lakes recharged by rivers flowing through deep canyons replaced the previously marine basins. During this time, the remaining bodies of water were either too salty or not salty enough for normal marine fauna to flourish. This was the so-called Messinian salinity crisis. Because Messinian sediments are essentially devoid of marine fossils this dramatic geological event has been difficult to date or place accurately in the global stratigraphic record, meaning in turn that its causes and environmental impact have been unclear.

Beginning in 1833 with Charles Lyell's

formal subdivision of geological strata into relative periods, based on the percentage of fossil shells representing living specimens found in the respective formations², marine fossils in sedimentary sequences have been the fundamental tool used to help impose chronological order on the stratigraphic record. Without marine fossils, the precise correlation of the Messinian salinity crisis with the global stratigraphic timescale has been an elusive goal. The significance of Messinian events can be judged from the estimate that as much as 6% of the ocean's salt was transformed into giant deposits left on the Mediterranean sea floor by evaporation³.

Now, the application of the relatively new stratigraphic technique of astrochronology, as reported by Krijgsman *et al.* on page 652 of this issue⁴, has enabled accurate dating of

these fossil-poor sediments. Alterations in Earth's obliquity and distance from the Sun occur in regular cycles, and the resulting changes in climate are reflected in the contents of sedimentary deposits. Astrochronology involves time-series analyses of a sedimentary sequence as related to these astronomically forced cycles. When supported by biochronology and magnetostratigraphy, the approach allows for an unprecedented precision in the definition of geological time, and consequently for an equally high-precision worldwide correlation of the Messinian events with other sedimentary sequences deposited during this period of Earth's history.

In the 1950s, geologists working in southern Europe were already beginning to recognize the immensity of the Messinian salinity crisis because of the unusual stratigraphic positioning of the evaporitic deposits, which are sandwiched between deep-water marine sediments, composed of clay and calcium carbonate, known as marls. In contrast to the salt minerals of the evaporites, which were precipitated from hypersaline waters, these marls represent deposition under normal open-marine conditions. The upper transition from continental evaporites to deep-sea sediments is particularly abrupt and can be explained only as the result of a huge flood in which Atlantic water poured across the Gibraltar arc and into the Mediterranean basin (Fig. 1).

This transition to full marine conditions was designated as the boundary between the Miocene and Pliocene epochs — which, dated at 5.33 million years (Myr) ago, is a major marker in the geological calendar⁵. The evaporitic deposits of the 'Gessoso-solfifera' formation in Sicily were taken as denoting the Messinian stage of geological time⁶. But defining these stratigraphic units or stratotype localities in Sicily, or for that matter anywhere else in the Mediterranean, violated long-held stratigraphic principles, because a continuous marine sequence straddling the Miocene–Pliocene boundary exists nowhere in the region. A stratigraphic controversy was thus born⁷.

That controversy, however, has not been limited to the problems of stratigraphic correlation. With the initiation of scientific ocean drilling, two expeditions of the Deep Sea Drilling Program (DSDP) in the Mediterranean showed that the Messinian evaporitic deposits were not restricted to the shallow marginal basins, where geologists had studied land sections. Astonishingly, they were widespread in all of the deeper Mediterranean basins, as imaged by basin-wide seismic profiles and sampled by drilling^{8,9}.

With the publication of these exciting results, debate over the Messinian salinity crisis erupted into a full-blown scientific fracas, with opposing camps developing

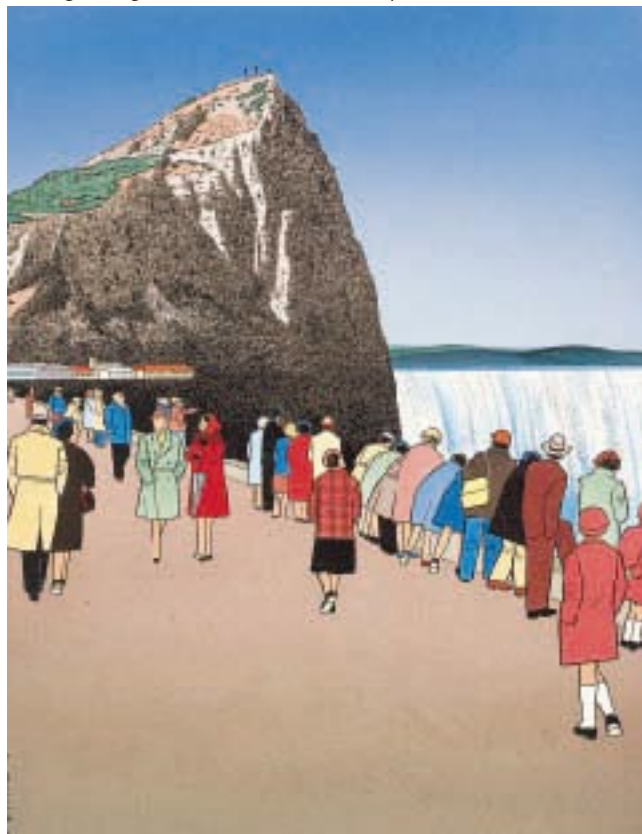


Figure 1 "The Filling of the Mediterranean Sea" by Guy Billout. This whimsical cartoon, first published in *The Atlantic Monthly* and mixing modern promenaders at the Rock of Gibraltar with an event that happened 5.33 million years ago, depicts the catastrophic flood that ended the Messinian salinity crisis when Atlantic waters re-entered the Mediterranean.

radically contrasting views and models about the environmental conditions surrounding the evaporite deposition⁷. Now, almost 30 years after the first DSDP discovery, the combatants may be able to reach an accord. Krijgsman and colleagues' application⁴ of astrochronology permits a new reading of the complex stratigraphic record produced by the Messinian events.

In my view, this work resolves several issues. First, as I have mentioned, the method sidesteps the need for marine fossils to date the Messinian sequences. Second, the new chronology shows that the beginning and end of the Messinian salinity crisis were synchronous throughout the Mediterranean; Krijgsman *et al.* date the beginning at 5.96 Myr and the end at 5.33 Myr ago. The causes and consequences of events can thus be distinguished. It is now clear that a tectonic closure in the general region of the modern Straits of Gibraltar, possibly augmented by a fall in sea level, led to the synchronous, basin-wide onset of evaporative conditions. The initial evaporative draw-down of water in the Mediterranean would have been gradual, with the development of both shallow-water and deep-water evaporitic sequences depending on the size and shape of individual basins. Many of the arguments put forward by the proponents of the various models can be accommodated within this picture.

Regardless of that, the massive flood that permanently ended evaporative conditions in the Mediterranean must have been caused by a dramatic rise in sea level. Such a rise would undoubtedly have had consequences elsewhere in the world. One example may have been the major flooding of the Bahamas Platform at around the Miocene–Pliocene boundary, which led to a huge 'back-

stepping' of the reef system to a higher, more central location on the platform¹⁰.

But what can explain the large erosional surfaces and deep canyons cut into the sea floor that are observed in parts of the Mediterranean^{8,9}? The extraordinary existence of very deep, dried-out basins during the Messinian remains an essential element of any such explanation, and our understanding will have a large gap in it as long as the complete sedimentary sequences in the deepest Mediterranean basins remain unexplored. It is only in the very deepest basins that sedimentation could have continued uninterrupted during the salinity crisis. Securing the necessary samples to fill in the missing period between 5.59 and 5.50 Myr ago will require advanced drilling technology to penetrate through the thick evaporite sequences to determine what lies below. Until this final 'Messinian gap' is closed, the giant salt deposits of the Messinian salinity crisis will remain a theme for speculation and argument. □

Judith A. McKenzie is at the Geological Institute, Swiss Federal Institute of Technology, ETH-Zentrum, CH-8092 Zürich, Switzerland.
e-mail: sediment@erdw.ethz.ch

1. Hsu, K. J. *The Mediterranean was a Desert* (Princeton Univ. Press, 1983).
2. Lyell, C. *Principles of Geology* (Murray, London, 1833).
3. Ryan, W. B. F. in *Messinian Events in the Mediterranean* (ed. Drogger, C. W.) 26–38 (K. Ned. Akad. Wetensch., Amsterdam, 1973).
4. Krijgsman, W., Hilgen, F. J., Raffi, I., Sierro, F. J. & Wilson, D. S. *Nature* **400**, 652–655 (1999).
5. Cita, M. B. *Micropaleontology Spec. Publ.* **1**, 1–30 (1975).
6. Selli, R. *Giorn. Geol.* **28**, 1–33 (1960).
7. Cita, M. B. in *Controversies in Modern Geology* (eds Müller, D. W., McKenzie, J. A. & Weissert, H.) 13–23 (Academic, London, 1991).
8. Ryan, W. B. F. *Init. Rep. DSDP Leg 13* (1973).
9. Hsu, K. J. *et al. Init. Rep. DSDP Leg 42A* (1978).
10. McKenzie, J. A., Spezzaferri, S. & Isern, A. *Mem. Soc. Geol. It.* **54**, 93–108 (1999).

components of the *V(D)J* recombination machinery are encoded by the recombination-activating genes, *RAG1* and *RAG2*. These genes are expressed in pro- and pre-lymphocytes, coincident with the two periods of *V(D)J* recombination. But if immature B cells in the bone marrow recognize a self antigen, they can maintain expression of *RAG1* and *RAG2*, and alter their BCR through a process known as receptor editing^{5,6}.

The RAG proteins initiate recombination by recognizing and cleaving chromosomal DNA at specific recombination sites. Mature B and T cells do not express the genes that encode these proteins — or so it was thought, until RAG messenger RNA transcripts, protein and *V(D)J* recombination activity were detected in B-lineage cells in the spleen, lymph nodes and tonsils^{7–13}. The two basic aspects of these findings were distinguished by their kinetics. First, within one to two days, cultured B cells from the spleen or lymph nodes could be stimulated (by cytokines^{9,11}, or by interaction with antigens¹³) to perform *de novo* immunoglobulin-gene recombination, replacing one BCR with another. Second, RAG-expressing B cells and evidence of recombination could be found late in the immune response (after 12–16 days) in splenic germinal centres^{7,10} — sites where mature B cells are activated by antigen and T cells to differentiate and secrete immunoglobulins. These results supported the conclusion that mature B cells can be induced by antigen to re-express *RAG1* and *RAG2*, and, by immunoglobulin-gene recombination, to generate new BCR specificities. With them, a central pillar of the clonal-selection theory seemed to collapse.

Much of this argument relied on the inference that expression of *RAG1* and *RAG2* is reactivated in mature B cells after having been turned off. But this had, in fact, never been demonstrated, because methods for detecting and quantifying expression of *RAG1* and *RAG2* in individual, live cells were not available. Yu *et al.*³ and Monroe *et al.*⁴ have now created strains of mice in which a green fluorescent protein (GFP) is a marker for cells that express *RAG2* (and, by inference, *RAG1*, because the two genes seem to be coordinately expressed).

Yu *et al.* used a large reporter transgene consisting of the RAG chromosomal region with GFP coding sequences substituted for those of *RAG2*. Monroe *et al.* created 'knock-in' mice, in which the endogenous *RAG2* locus directs expression of a *RAG2*–GFP fusion protein. Both approaches yield GFP expression in the expected subsets of developing B and T cells, but each has a limitation. Although the GFP protein encoded by Yu and colleagues' transgene is highly expressed, it seems to be more stable than endogenous *RAG2* protein. The result is some 'green' cells containing little or no

Immunology

Developing B-cell theories

David G. Schatz

Mark Twain once wrote, "The reports of my death are greatly exaggerated". Perhaps the same can now be said for some aspects of 'clonal selection' in the adaptive immune response^{1,2}. According to this theory, B and T cells have a diverse array of antigen receptors before they encounter antigens (bacteria or viruses, for example). Those cells whose receptors recognize antigen are then selected to proliferate and differentiate into effector and memory cells.

The possibility that antigen-specific interactions trigger cells to produce a new array of receptors is specifically forbidden by the clonal-selection theory. Yet a series of surprising studies has implied that, in mature B cells, assembly of antigen-receptor genes can be reactivated in response to anti-

gen. Now, reports by Yu *et al.*³ (on page 682 of this issue) and Monroe *et al.*⁴ (in *Immunity*) describe a powerful new tool that should allow a more sophisticated analysis of B- and T-cell development, and indicate how the previous findings might be reconciled with the clonal-selection theory.

The development of B and T cells is organized around *V(D)J* recombination — the assembly of immunoglobulin and T-cell receptor (TCR) genes from component *V* (variable), *D* (diversity) and *J* (joining) gene segments. Two periods of *V(D)J* recombination, at the pro- and pre-lymphocyte stages of development, lead to expression of the B-cell receptor (BCR) and TCR on the surface of immature B and T cells, respectively (Fig. 1, overleaf). The essential, lymphocyte-specific

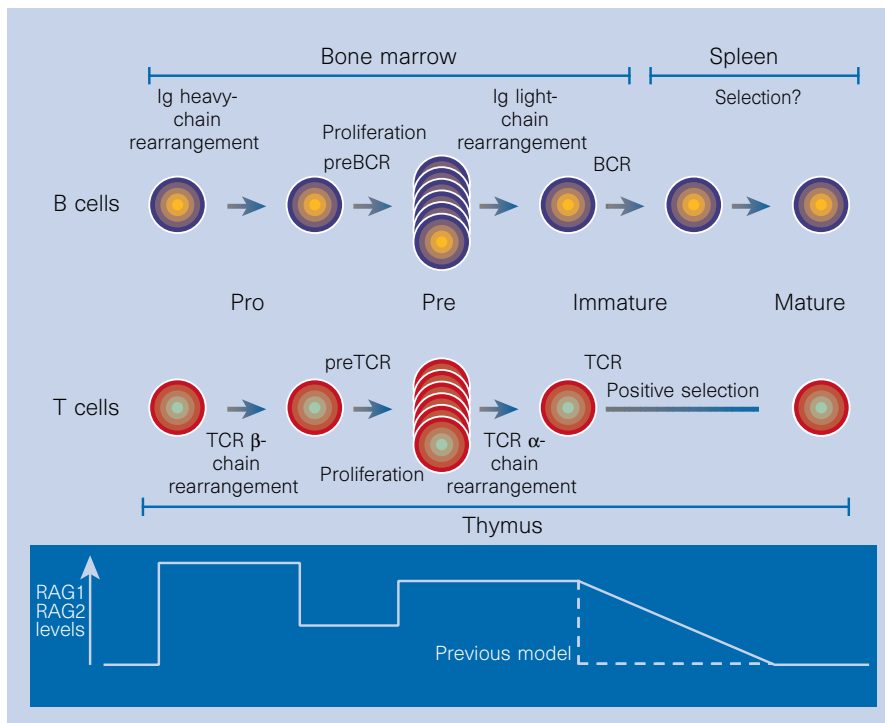


Figure 1 B- and T-cell development. Immunoglobulin (Ig) heavy-chain and T-cell receptor (TCR) β -chain genes are assembled by $V(D)J$ recombination in pro-B and pro-T cells, respectively. Heavy and β chains then pair with 'surrogate' receptor subunits to generate the pre-B-cell receptor (pre-BCR) and pre-TCR, respectively. These receptors are critical for developmental and proliferative events leading to pre-B and pre-T cells. Here, a second wave of recombination occurs, assembling immunoglobulin light-chain and TCR α -chain genes, and allow expression of the complete BCR and TCR. Expression of the *RAG1* and *RAG2* genes continues in TCR-expressing thymocytes until the cells either receive a positive selection signal or they die. Expression of the *RAG* genes was previously thought to be terminated in immature B cells (dotted line), but the results of Yu *et al.*³ and Monroe *et al.*⁴ indicate that expression drops gradually as the cells develop further.

RAG2 protein. In contrast, Monroe and colleagues' *RAG2*-GFP protein appears to be appropriately regulated. But it could conceivably be relatively unstable, and, because it is expressed at only modest levels, detection of cells with low expression of the gene is a problem.

Both groups detected a subset of *RAG*-expressing B cells in the spleen. Expression of the *RAG* genes was previously thought to terminate in the bone marrow, but it now seems that expression decreases relatively slowly as B cells mature, and continues in at least some splenic B cells (Fig. 1). Yu and colleagues also show that adoptively transferred mature splenic B cells that do not express the *RAG* genes do not reactivate expression of GFP or *RAG* when stimulated by antigen or cytokines. How, then, can one explain earlier results showing rapid induction of $V(D)J$ recombination in splenic and lymph-node B cells? Yu *et al.* propose that, when a developing B cell encounters a low-affinity antigen, the maturation process is slowed down, prolonging the time that the B cell expresses the *RAG* genes. This implies that rapidly induced recombination occurs in immature B cells which have never stopped expressing *RAG*, and that it occurs without an increase in *RAG* expression lev-

els. It is not clear how to reconcile this idea with the increased *RAG* expression triggered by cytokines or antigen in B cells in the lymph nodes^{8,14}. Nor is it easy to explain the extensive $V(D)J$ recombination in stimulated splenic B cells if, indeed, they continue to express the very low levels of *RAG1* and *RAG2* seen in unstimulated cells.

Perhaps the greater mystery is how to explain the delayed appearance of *RAG*-expressing B cells in splenic germinal centres^{7,10}. Monroe *et al.*⁴ show that such cells have many phenotypic properties of pre-B cells, and these authors advance some intriguing theories to account for this. One possibility is that, during an immune response, some mature B cells reacquire immature characteristics (including expression of *RAG*), as originally proposed by Han *et al.*⁷. But Yu and colleagues' results argue against this idea. Alternatively, these mysterious cells may be true pre-B cells that either arose in the spleen or were recruited from the bone marrow. Recruitment of pre-B cells to sites of an active immune response would be a new and unexpected phenomenon.

Finally, Yu *et al.*³ and Monroe *et al.*⁴ extend the striking parallels between B- and T-cell development by showing that, in both lineages, termination of *RAG* expression



100 YEARS AGO

It should not be necessary at this time of day to emphasise the fact of the imperial character of the Royal Gardens, Kew, still it would appear there are many inhabitants of Great Britain whose notion of the value of this establishment is limited by their desire for a local public park suited to the recreation of dwellers in and about London. Several incidents have of late shown this – witness the recent preposterous proposal brought forward in the House of Commons to throw the gardens open to cyclists! Suggestions of this kind are on the face of them, to those aware of the true character of the gardens, too absurd for discussion, yet there is an element of danger in this appeal to the selfish instincts of that large body of pleasure-seekers who are veritable Gallios in their contempt for science, especially when its just claims place an obstacle to the gratification of their pleasure whims. From *Nature* 10 August 1899.

50 YEARS AGO

Now that the long-established mechanical unit, the joule, is to assume so much greater importance, there is one matter ... that should be decided, and that is how the word is to be pronounced. There was not long ago a correspondence on the subject in *Nature* which revealed the most surprising disagreement, and more recent inquiries among various people in Manchester confirm these doubts. Some make the great man's name rhyme with 'cool', some with 'cowl', and some even with 'coal'. Thus, Sir Arthur Schuster called him 'Jowl', while Prof. H. B. Dixon, who had a curiosity in such matters, said the name was disyllabic as 'Jo-ull', which would be nearly 'Joal'. On the other hand, Osborne Reynolds in his memoir on him, though he never states how the name is spoken, mentions that it derives from the village of Youlgreave in Derbyshire from which his family came, which would make it 'Jool', and this is the preference of at least some of the surviving family. However all this may be, the word is now to be international, and it is unreasonable to make foreigners suffer from the horrors of English orthography; it therefore seems best that the joule as a unit of heat should rhyme with the word 'cool' – of course, giving the initial 'j' its proper English value.

From *Nature* 13 August 1949.

occurs *after* the cells express antigen receptor on their surface (Fig. 1). So, how are RAG expression and V(D)J recombination terminated in immature cells? In the thymus, immature T cells whose TCRs are weakly reactive to self major-histocompatibility antigens receive a 'positive selection' signal. This triggers downregulation of RAG expression and maturation. By analogy, it is tempting to think that immature B cells are positively selected based on the specificity of the BCR. Indeed, strong engagement of the BCR on such cells can downregulate RAG expression and V(D)J recombination^{12,13,15}. If B cells do turn out to be positively selected, the next steps will be to work out how such selection shapes the BCR repertoire, and what the selecting ligands are. □

David G. Schatz is at the Howard Hughes Medical Institute, Section of Immunobiology, Yale University

School of Medicine, 310 Cedar Street, New Haven, Connecticut 06520-8011, USA.

e-mail: david.schatz@yale.edu

1. Burnet, F. M. *The Clonal Selection Theory of Acquired Immunity* (Cambridge Univ. Press, 1959).
2. Talmage, D. W. *Science* **129**, 1643–1648 (1959).
3. Yu, W. *et al. Nature* **400**, 682–687 (1999).
4. Monroe, R. J. *et al. Immunity* (in the press).
5. Tiegs, S. L., Russell, D. M. & Nemazee, D. J. *Exp. Med.* **177**, 1009–1020 (1993).
6. Gay, D., Saunders, T., Camper, S. & Weigert, M. J. *Exp. Med.* **177**, 999–1008 (1993).
7. Han, S., Zheng, B., Schatz, D. G., Spanopoulou, E. & Kelsoe, G. *Science* **274**, 2094–2097 (1996).
8. Hikida, M. *et al. Science* **274**, 2092–2094 (1996).
9. Papavasiliou, F. *et al. Science* **278**, 298–301 (1997).
10. Han, S. *et al. Science* **278**, 301–305 (1997).
11. Hikida, M. & Ohmori, H. *J. Exp. Med.* **187**, 795–799 (1998).
12. Meffre, E. *et al. J. Exp. Med.* **188**, 765–772 (1998).
13. Hertz, M., Kouskoff, V., Nakamura, T. & Nemazee, D. *Nature* **394**, 292–295 (1998).
14. Hikida, M., Mori, M., Kawabata, T., Takai, T. & Ohmori, H. *J. Immunol.* **158**, 2509–2512 (1997).
15. Ma, A. *et al. EMBO J.* **11**, 2727–2734 (1992).

Crystal structures

Tiling by numbers

Michael O'Keeffe

There is a long history of overlapping problems in chemistry and geometry. A classic example is the problem of the densest packing of equal spheres. Johannes Kepler conjectured in 1611 that the face-centred cubic lattice ('cubic closest packing') was the solution. In 1883, W. Barlow showed that there is an infinite family of equally dense sphere packings, of which the hexagonal closest packing is the simplest member, and correctly suggested that metallic elements would have such structures. Later, Barlow and Sir William Pope proposed structures based on sphere packings, correctly again, for simple salts. This was exactly what W. H. and W. L. Bragg needed for the first X-ray diffraction solutions of crystal structures, such as those of ZnS and NaCl in 1913.

Nowadays, chemists rely heavily on theoretically derived structures to help determine the structures of new materials, such as zeolites, which are based on low-density atomic packings. For years, the goal of enumerating possible structure types has absorbed the efforts of many crystal chemists, and until now this work has been empirical and intuitive. But for the first time, in a study¹ described by Delgado Friedrichs *et al.* on page 644 of this issue, a mathematically rigorous and complete enumeration of the main classes of such structures is achieved. This work is of considerable importance to both chemistry and mathematics.

Filling space with polyhedra is a rich source of mathematical problems, many of them still unsolved. A classic problem asks for the different types of convex polyhedra that fill space when repeated only by translation. In 1885, E. S. Federov showed that there are only five types. In one of these structures,

four polyhedra meet at each vertex, three at each edge and two at each face. The set of vertices and edges form a network in which every vertex is joined to four others by edges, creating a structure that chemists call a 'four-connected net'. In its most symmetrical form, all vertices are related by symmetry (Fig. 1). This 'uninodal' structure has been discussed extensively, and Lord Kelvin suggested that with suitably curved edges it would be the lowest-energy structure (that is, lowest surface area) of monodisperse foam. In fact, the vertices and edges of any stable foam form a four-connected net. But in 1994 it was shown² that if the foam bubbles

are not required to be congruent, the Kelvin conjecture is false. A structure with lower energy was found that was based on a packing of two kinds of polyhedra, also known as the type I clathrate hydrate crystal structure — another illustration of the interplay between chemistry and mathematics.

In crystal chemistry the Kelvin structure is better known as the structure of the mineral sodalite. In the 'sodalite structure', T atoms (Si or Al) are surrounded by O atoms at the vertices of TO₄ tetrahedra, which share vertices to form a TO₂ framework. So the T atoms are at the vertices of the net and T–O–T links form the edges. Other tetrahedral frameworks form the structures of the most common minerals (feldspar and quartz) and of solid and liquid water.

Synthetic zeolites with tetrahedral structures are some of the most valuable industrial chemicals, such as highly specific catalysts in petrochemical reactions and components of detergents. Enormous effort has been devoted to discovering new materials. A key structure is that of the mineral faujasite, which is also based on a uninodal four-connected net derived from a polyhedral packing, but here there are three different kinds of polyhedra involved. What other structures of this type are possible? Delgado Friedrichs *et al.*¹ now show that there are exactly nine uninodal, 117 binodal and 926 trinodal structures of this type derived from polyhedral packings.

There are two practical consequences of this work. First, increasingly sophisticated understanding of synthetic methods makes rational synthesis directed at a specific target structure possible, and clearly it is of value to know all the possibilities. Second, determining the crystal structure of new materials that are not available as single crystals is not easy, and in many cases this can be solved only

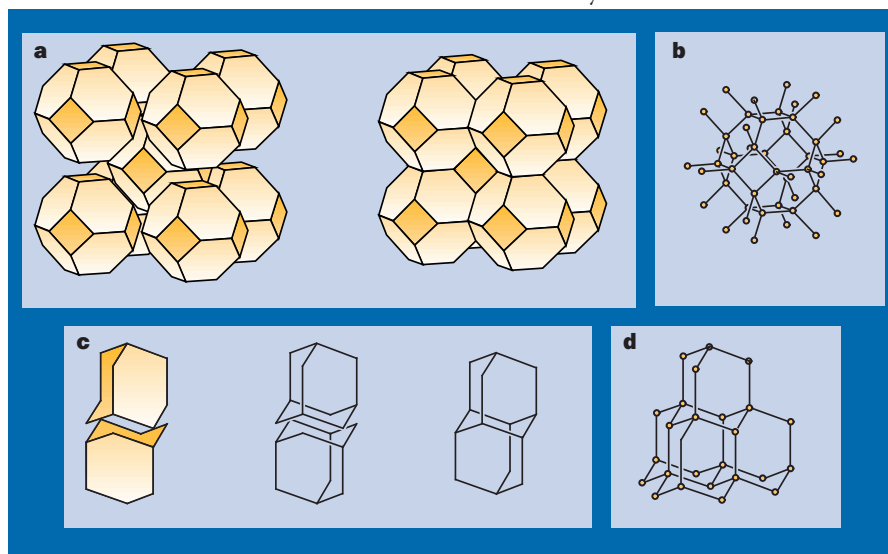


Figure 1 Space-filling tiles and networks. a, Tiles showing the truncated octahedra of the Kelvin (sodalite) structure, and b, A fragment of the Kelvin four-connected net. c, Two of the tiles of the diamond structure known as saddle polyhedra, and d, A fragment of the diamond four-connected net. Systematic enumeration of all such four-connected nets by Delgado Friedrichs *et al.*¹ should lead to new structures for zeolites.

with the help of theoretical trial structures³.

The authors' enumeration of structures proceeds from the point of view of the vertices; that is, first uninodal structures, then binodal, and so on, are enumerated. Nature also works that way: with some exceptions, the most common and important structures have a relatively small number of vertex types. In contrast, if one tried to evaluate four-connected nets with only one kind of tile, the number of possibilities would be infinite and the structures would rapidly become complex and unlike real crystal structures⁴. It is worth noting that the question of whether the tiles are convex or not does not arise in the new work, whereas most other mathematical results are confined to tilings by convex polyhedra⁵.

There is particular interest in four-connected nets generated by packing polyhedra with only five- and six-sided faces⁶. The dual structures obtained by placing vertices in the centres of the polyhedra are dense sphere packings involving more than one kind of sphere, and include the most common intermetallic structure types. The simplest of these (known to chemists as MgCu₂ and Cr₃Si) are formed from dual trinodal nets (one of which corresponds to the lowest-energy foam), and appear in Delgado Friedrichs and colleagues' study. It is reassuring to know that there are no other simpler structures that have been missed. The possibility of systematically enumerating packings with more than one kind of sphere may lead to identification of new intermetallic structure types.

Not all four-connected nets are derived from packing conventional polyhedra. Indeed, the simplest net is diamond (Fig. 1), which is made up of space-filling tiles with curved faces, in which only two edges meet at some vertices (these have been called saddle polyhedra). Delgado Friedrichs *et al.* have also enumerated the structures based on these 'quasi-simple' tiles. The packings are generally denser, and not so interesting as potential zeolites, but are valuable as possible structures of high-pressure silicas and ices.

These methods have solved another longstanding mathematical problem⁷. The challenge relates to tiling space by congruent tetrahedra (curiously, it was thought for many centuries that only regular tetrahedra could tile space, although an elementary calculation of the angle between the faces would show that is impossible). It turns out that there are only nine topological types of tetrahedral tiling. As each symmetry-related tetrahedron shares a face with four others, the centres of the tetrahedra form a four-connected net that is one of nine uninodal nets derived from simple tilings¹.

In one of these tetrahedral packings (corresponding to the sodalite or Kelvin structure) all the vertices of the tetrahedra are congruent, and in fact correspond to the body-centred cubic (b.c.c.) sphere packing.

So the b.c.c. structure is the only packing of equivalent (symmetry-related) spheres that divides space into equivalent tetrahedra (of course it has many other special properties⁸). From the point of view of simplicity, it is satisfying to know that the b.c.c. lattice is the most common structure for elements (such as iron at normal temperatures and pressures) that do not adopt a Barlow packing as crystal structure. □

Michael O'Keefe is in the Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604, USA.

e-mail: mokeeffe@asu.edu

1. Delgado Friedrichs, O., Dress, A. W. M., Huson, D. H., Klinowski, J. & Mackay, A. L. *Nature* **400**, 644–647 (1999).
2. Weaire, D. & Phelan, R. *Phil. Mag. Letts.* **69**, 107–110 (1994).
3. Newsam, J. M., Treacy, M. M. J., Koetsier, W. T. & de Gruyter, C. B. *Proc. R. Soc. Lond. A* **420**, 375–405 (1988).
4. O'Keefe, M. Z. *Kristallogr.* **214**, 438–443 (1999).
5. Grünbaum, B. & Shephard, G. C. *Bull. Am. Math. Soc.* **3**, 951–972 (1980).
6. O'Keefe, M. in *Foams and Emulsions* (eds Sadoc, J. F. & Rivier, N.) 403–421 (Kluwer, Dordrecht, 1999).
7. Delgado Friedrichs, O. & Huson, D. H. *Discr. Comput. Geom.* **21**, 299–315 (1999).
8. Conway, J. H. & Sloane, N. J. A. *Sphere Packings, Lattices and Groups* 3rd edn (Springer, New York, 1999).

Evolution

Creatures from another world

Inman Harvey

Genetic experiments have been carried out on an organism that shares no common ancestry with any other terrestrial life-form. The experiments, reported by Lenski and colleagues on page 661 of this issue¹, were neither *in vivo* nor *in vitro*; they used digital organisms created in a computer — '*in silico*'. The results, claim the authors, support the view that complex interactions between multiple mutations are a general feature of genetic systems. In living organisms these interactions are difficult to measure, but they could be related to such thorny evolutionary issues as the origin and maintenance of sex. Can these findings from a digital world be extrapolated to the real world?

The study of artificial life springs largely from the development of computers as modelling tools, and from a series of conferences initiated by Langton² in 1987. The field brings together computer scientists and biologists with disparate interests; it bears the same sort of relationship to biology as artificial intelligence bears to cognitive science. This means that some people in the field want to pick up tips from biologists in order to build technological artefacts — perhaps robots, or software — with some of the adaptive, evolved properties of natural living organisms. Others are seeking inspiration in the other direction, using artefacts created in computer programs as a new tool to investigate questions in theoretical biology; the paper by Lenski and colleagues¹ falls into this category, aiming to investigate the effects of genetic mutations on the fitness of digital organisms.

Even in the earliest days of computing, Barricelli³ was exploring evolutionary phenomena using artificial self-reproducing entities within computer programs. More recently, Ray stimulated interest in this field by generating an evolving 'ecosystem' within a computer that demonstrated parasitic and hyper-parasitic behaviour^{4,5}. The work by Lenski *et al.* uses digital organisms distantly related to Ray's. The organisms are specified by strings of sequential primitive computer

instructions that can be thought of as strings of artificial DNA. This mixing of ideas from biology and computer science can lead to culture clashes. For example, the computer scientist may think in terms of such a genome being a vector specifying a point in some high-dimensional search space.

An artificial world is constructed in computer memory, within which several of these digital organisms can replicate and mutate. Selection between them arises from competition for limited 'energy' — processing time from the computer CPU. Once heredity, variation and selection have been set up in this digital world, Darwinian evolution takes over within the computer memory, and the experimenters become observers. The dynamics of evolutionary processes are so complex that one cannot predict the results before running the experiments.

The experimenters can, however, make changes in parameters and note the resulting changes; and they have the enormous advantage over the conventional biologist that all relevant data can be recorded without fear of error. Lenski *et al.* manipulated the selective pressures so that both 'simple' and 'complex' digital organisms evolved. Simple organisms were under selective pressure only to replicate as fast as possible. Complex organisms, on the other hand, were also 'rewarded' (with extra CPU time) for performing certain mathematical functions. Lenski and colleagues then measured the fitnesses (replicative rates) of many mutant copies of these evolved organisms. Mutations are usually harmful, and each successive mutation tends to reduce fitness further. In the simplest, multiplicative situation, a mutation has the same relative effect on the fitness of an unmutated organism as on one that is already mutated. There can, however, be different kinds of interaction between successive mutations: synergistic epistasis, where the cumulative effects of multiple mutations are worse than is implied by their individual effects, and antagonistic epistasis, where the effects are less damaging.

In these experiments the complex digital organisms showed antagonistic epistasis; they were relatively robust to the effect of multiple mutations. When the overall results were studied, simple organisms appeared to display no significant epistasis. The ease of data collection in such computer experiments, however, allowed exhaustive analysis. This showed that there were indeed some epistatic interactions in these simple organisms, but that the synergistic and antagonistic effects largely balanced out.

Do these results from an artificial world have any relevance to organisms in the natural world? Workers in this field argue that these techniques allow one to make generalizations about evolving systems, stepping out of the restrictions of a terrestrial, DNA-based world, where all are derived from a common ancestor — but care is needed. To be able to make such generalizations, the conditions under which these digital organisms develop must be set up without any bias favouring one result or another, which requires great care.

The approach used by Lenski and colleagues¹ raises more questions than do their specific findings. When theoretical biolo-

gists reduce a complex ecosystem to a simple analytical formula, it is clear that that they are assuming, for the sake of their theory, that only a small number of factors are relevant. The Artificial Life approach comes into its own when analysis is difficult but computer simulations feasible; however, the assumptions used may be less obvious. The advantages of experiments using digital organisms suggest that they will find further uses in those vast areas of evolutionary theory where analysis is intractable. However, considerable debate can be expected before a consensus is reached on just what is necessary for results from a synthesized world to be seen as relevant to the natural world. □

Inman Harvey is at the Centre for the Study of Evolution, and the Centre for Computational Neuroscience and Robotics, University of Sussex, Brighton BN1 9QH, UK.

e-mail: inmanh@cogs.susx.ac.uk

1. Lenski, R. E., Ofria, C., Collier, T. C. & Adami, C. *Nature* **400**, 661–664 (1999).
2. Langton, C. (ed.) *Artificial Life* (Addison-Wesley, Redwood City, California, 1989).
3. Barricelli, N. A. *Acta Biotheoretica* **XVI**, 69–98 (1962).
4. Ray, T. S. in *Artificial Life II* (eds Langton, C. G., Taylor, C., Farmer, J. D. & Rasmussen, S.) 371–408 (Addison-Wesley, Redwood City, California, 1991).
5. Maynard Smith, J. *Nature* **335**, 772–773 (1992).

Fuel-cell technology

Running on natural gas

Brian C. H. Steele

Deregulation of the electric power industry and more stringent emission controls are stimulating investment into fuel-cell systems, which are virtually free from noxious emissions of nitrogen oxides normally associated with burning fossil fuels. They can convert chemical energy directly into electricity with greater efficiency than most other devices, which both conserves fuel resources and reduces CO₂ emissions. Although hydrogen is considered the ideal fuel for many energy-conversion systems, its widespread use is dependent on technological breakthroughs in its cost and storage. So it is best to assume that in the immediate future, except for a few niche markets, fuel cells will have to use hydrocarbons (such as methane) or alcohols (such as methanol) as fuel. In this context, it is timely to examine the merits of direct electrochemical oxidation of methane in a fuel cell reported by Murray *et al.*¹ on page 649 of this issue.

Using methane as the primary fuel favours devices that operate at high temperatures², such as molten carbonate and solid-oxide fuel cells (SOFC). These are able to electrochemically oxidize a mixture of hydrogen and carbon monoxide, which is generated internally from the natural gas (methane) within the fuel stack. The development of SOFC technology has been principally directed towards large-scale power

generation, and early commercial devices operating on natural gas at around 900 °C are expected to produce 1–3 megawatts (MW) of power³. Such fuel cells use anodes that contain nickel to catalyse the conversion of methane to H₂ and CO, but large amounts of

steam have to be added to prevent carbon deposition on the nickel-ceramic anodes, which severely degrades their performance. The steam-conversion reaction is endothermic and has to be sustained by significant thermal energy, which is available in the high-temperature SOFCs. In contrast, low-temperature fuel cells need an external processor to produce H₂ and remove CO, which not only reduces the efficiency of the system but also increases its complexity and cost (Fig. 1).

Over the past decade there has been a growing realization that there could be an important market for small (1–10 kW) thick-film SOFC stacks operating at intermediate temperatures (500–700 °C). These systems could provide heating and electricity for buildings⁴ (domestic and small commercial units), or provide electrical power for auxiliary functions (such as electric windows or air conditioning) in vehicles⁵. Although the provision of excess steam is appropriate for relatively large (> 1 MW) SOFC systems, it introduces cost and efficiency penalties as the system size is reduced. Direct electrochemical oxidation of dry methane in intermediate-temperature SOFCs, as developed by Murray *et al.*¹, would have many advantages in small-scale applications.

Methane oxidation and activation has been investigated with a wide range of oxide catalysts^{6,7}, but at present no consensus exists about the mechanisms involved. It has been suggested⁸ that oxygen surface species (O[−], O^{2−}, O₂[−] and so on) and lattice oxygen in the electrodes are responsible for different types of behaviour, and that high catalytic activity is often associated with mixed electronic and oxygen-ion conductivity.

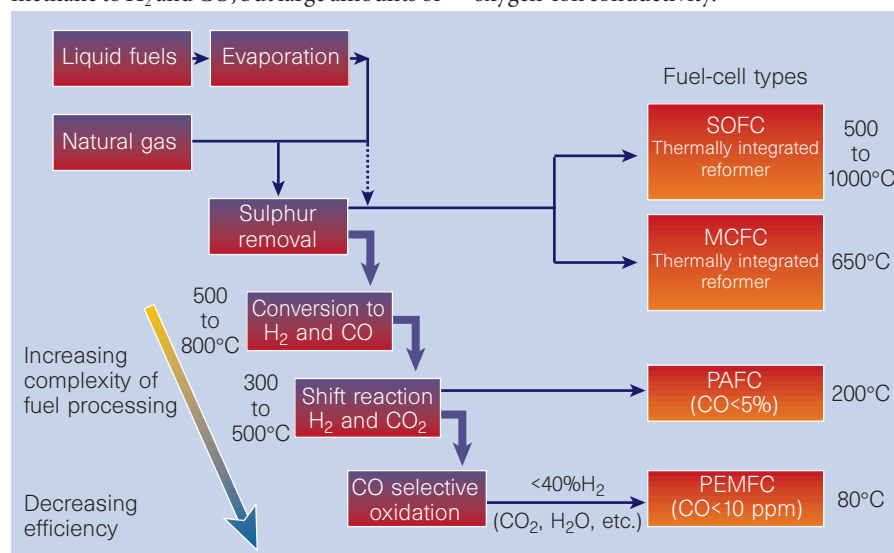


Figure 1 Fuel processing reactions (reforming of methane to hydrogen fuel) and their influence on the complexity and efficiency of different types of fuel cell. The intermediate-temperature solid-oxide fuel cells (SOFCs) developed by Murray *et al.*¹ bypass the need for methane steam reforming by directly oxidizing methane to CO₂ and H₂O. Devices that use direct methane oxidation should find new applications in areas where small-scale units are appropriate. MCFC, molten-carbonate fuel cell; PAFC, phosphoric-acid fuel cell; PEMFC, proton-exchange-membrane fuel cell.

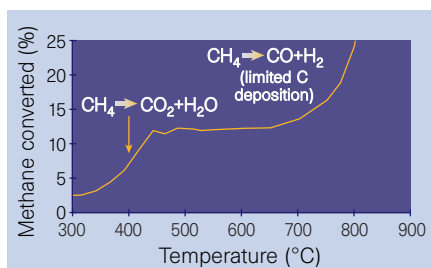


Figure 2 Conversion of methane by an oxide catalyst⁷. The percentage conversion of CH₄ by La_{0.8}Ca_{0.2}CrO₃ powder is shown as a function of temperature when fed with a mixture of methane and oxygen (diluted in helium) in the ratio 5:1 (by volume). Methane is directly oxidized to CO₂ and H₂O at temperatures around 400 °C. At higher temperatures CH₄ is reformed to H₂ and CO fuel with minor carbon deposition on the oxide catalyst.

In the SOFC research community, doped La_{0.8}Ca_{0.2}CrO₃ (LCC) and doped Ce_{0.9}Gd_{0.1}O_{1.95} (CGO) have received the most attention. LCC is a promising anode material^{7,9} because of its wide thermodynamic stability range, and existing use as an electrical connector between single fuel cells. Catalysts that contain ceria (such as CGO) have attracted interest¹⁰ because it has been known for decades¹¹ that the use of doped ceria can improve the performance of SOFC anodes. Both LCC and CGO have good catalytic activity for methane oxidation. Steady-state measurements⁷ with a limited supply of oxygen gas indicate that LCC oxidizes methane to CO₂ and H₂O at around 400 °C (Fig. 2). The reaction with CGO occurs at lower temperatures, around 300 °C. Exposure to residual excess methane confirms that only small amounts of carbon are deposited on both oxides at high temperatures (800–900 °C). Moreover, if carbon is intentionally deposited on the oxides it can be removed¹² at relatively low temperatures (300–400 °C) by the introduction of oxygen. The excellent catalytic behaviour of the ceria-based materials is not totally unexpected because CeO₂–ZrO₂ (CZO) catalysts are used in catalytic converters fitted to car-exhaust systems. It appears that CGO and CZO promote the reduction of Ce⁴⁺ to Ce³⁺, which enhances oxygen-exchange processes¹³ and associated catalytic reactions.

These catalytic studies suggest that the use of oxide electrodes instead of the standard composite nickel-ceramic anodes could have advantages in producing direct electrochemical oxidation of dry methane. But so far it has been impossible to synthesize a single-phase oxide material that satisfies all the criteria specified¹⁴ for use at intermediate temperatures.

The experiments of Murray *et al.*¹ use electrodes consisting of a thin layer (0.5 μm) of yttria-doped ceria (YDC) together with

a thicker layer (2 μm) of a nickel-ceramic composite as the current collector, demonstrating one way to overcome the relatively poor electronic conductivity of YDC. Yet although it works in the laboratory, this solution is unlikely to be adopted by SOFC developers for two reasons. First, natural gas usually contains higher alkanes (C₂H₆, C₃H₈ and so on) that are more easily dissociated than methane, resulting in carbon deposition on nickel at lower temperatures. Second, a number of investigations with intermediate-temperature SOFC stacks has revealed that when nearly all the fuel is converted to steam and CO₂ the nickel can be oxidized to NiO with a concomitant degradation in the anode performance.

A more promising strategy is to fabricate a composite anode in which doped CeO₂ is mixed with another oxide exhibiting good electronic conductivity, so avoiding the use of nickel altogether. This approach is being examined by a number of laboratories around the world and, if successful, is likely to speed up the commercialization of intermediate-temperature SOFC for small-scale applications. □

Brian C. H. Steele is in the Department of Materials, Prince Consort Road, Imperial College, London SW7 2AZ, UK.
e-mail: b.steele@ic.ac.uk

1. Murray, E. P., Tsai, T. & Barnett, S. A. *Nature* **400**, 649–651 (1999).
2. Steele, B. C. H. *CR Acad. Sci. I, Serie IIC*, 533–543 (1998).
3. Casanova, A. J. *Power Sources* **71**, 65–70 (1998).
4. Diethelm, R. *et al.* in *Proc. 3rd European SOFC Forum* (ed. Stevens, P.) 87–93 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 1998).
5. Morton, I. in *Automotive Engineer* June (1999).
6. Doshi, R., Alcock, C. B., Gunasekaran, N. & Carberry, J. J. *J. Catalysis* **140**, 557–563 (1993).
7. Middleton, P. H. *et al.* in *Solid Oxide Fuel Cells III Vols 93–94* (eds Singhal, S. & Iwahara, H.) 542–551 (Electrochem. Soc., Pennington, New Jersey, 1993).
8. Gellings, P. J. & Bouwmeester, H. J. M. *Catalysis Today* **12**, 1–105 (1992).
9. Steir, J., van Herle, J. & McEvoy, A. J. in *Proc. 3rd European SOFC Forum* (ed. Stevens, P.) 267–276 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 1998).
10. Marina, O. A., Bagger, C., Primdahl, S. & Mogensen, M. *Solid State Ionics* **123**, 199–208 (1999).
11. Takahashi, T. in *Physics of Electrolytes Vol. 2* (ed. Hladik, J.) 989–1049 (Academic, London, 1972).
12. Ramirez-Cabrera, E., Atkinson, A. & Chadwick, D. *Solid State Ionics* Abstr. I-10 (in the press).
13. Naito, H., Sakai, N., Otake, T., Yugami, H. & Yokawa, H. *Solid State Ionics* Abstr. C-07 (in the press).
14. Steele, B. C. H., Middleton, P. H. & Rudkin, R. A. *Solid State Ionics* **40/41**, 388–393 (1990).

Neurobiology

Oscillations in the basal ganglia

Thomas Wichmann and Mahlon R. DeLong

The basal ganglia constitute a network of brain structures that are components of larger loops connecting cortical and subcortical structures. Like many neuronal assemblies in the mammalian nervous system, this network seems to have two distinct operational modes. Classic studies of neuronal activity in the basal ganglia have focused on the single-spike mode, where neuronal discharge reflects information transfer or processing. But it has become clear that basal ganglia neurons can also discharge in a rhythmic, oscillatory-burst mode, in which individual bursts are unrelated to external events. As suggested for other systems that involve burst discharges, these bursts may signal and facilitate changes in functional states of the nervous system, or they could provide a representation of time^{1,2}. But what is the origin of these oscillatory bursts? On page 677 of this issue, Plenz and Kitai³ show that, in a cell-culture environment, synchronized oscillatory-burst discharges in one of the main basal ganglia structures, the subthalamic nucleus (STN), may depend on anatomical projections from another — the external segment of the globus pallidus (GPe; Fig. 1, overleaf).

Oscillatory discharge in subcortical regions, conveyed from the thalamus to the cortex via the thalamo-cortical network, may have powerful effects on cortical processing. To date, these changes have been studied

mainly in areas of the thalamus concerned with sensory processing. For example, under conditions of little or no sensory input — such as during non-REM sleep — the oscillatory activity of the thalamo-cortical system is dominated by synchronized, low-frequency oscillations generated in the thalamus². Waking is associated with desynchronization of thalamo-cortical activity, unmasking the less synchronized, higher-frequency⁴ rhythms that prevail within the cortex during wakefulness, and allowing faithful transfer of information to the cortex. Changes of cortical function on a smaller scale (such as the changes that underlie selective attention⁴) could be mediated by precisely controlling the degree of synchronized oscillatory activity along specific parts of the thalamo-cortical network. The time-signalling aspects of the oscillatory activity may help to coordinate activity patterns in spatially separate regions of cortex⁵. In this way, for instance, sensory features detected by spatially segregated areas of the cortex could be integrated perceptually (a process known as 'binding')⁶.

In all current models of basal ganglia function, the STN drives activity in the basal ganglia output nuclei — the internal segment of the globus pallidus (GPi) and the pars reticulata of the substantia nigra (SNr) — and is a port of entry to the basal ganglia for cortical inputs (Fig. 1). Activity changes in the STN strongly affect the average rates and

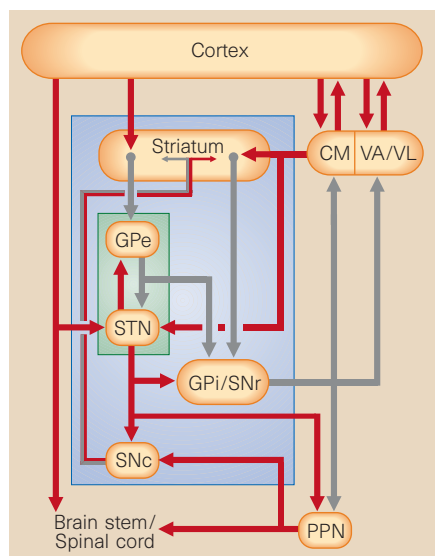


Figure 1 The basal ganglia/thalamo-cortical circuitry (simplified). The basal ganglia (blue box) consist of the striatum, the external and internal segments of the globus pallidus (GPe and GPi, respectively), the pars compacta and pars reticulata of the substantia nigra (SNc and SNr, respectively), and the subthalamic nucleus (STN). These structures are linked via a complex network of excitatory and inhibitory connections (shown as red and grey arrows, respectively). Their external connections include cortical areas, the thalamic centromedian (CM), ventral anterior (VA) and ventral lateral (VL) nuclei, and the pedunculopontine nucleus (PPN). The GPe/STN 'pacemaker' circuitry proposed by Plenz and Kitai³ is shown inside the green box.

occurrence of non-oscillatory bursts in the GPe and GPi⁷. Plenz and Kitai³ now show that the STN — and its link with the GPe — may also be important in generating synchronized oscillatory discharge in the basal ganglia. Previous brain-slice-recording studies have emphasized that individual neurons in the STN and GPe can generate oscillatory discharge through their intrinsic membrane properties^{8,9}. In the STN, this occurs particularly during hyperpolarization⁸. Plenz and Kitai provide evidence that the STN/GPe circuitry may generate the necessary conditions for such burst discharges to occur and be maintained in synchrony.

Within the basal ganglia/thalamo-cortical network, oscillatory discharge has been directly recorded in corticostriatal neurons¹⁰, tonically active striatal interneurons^{11,12}, the substantia nigra pars compacta (SNc)¹³, both pallidal segments and the STN^{8,14} (Fig. 1). Hypotheses about the physiological importance of oscillatory discharge in the basal ganglia are tentative at best. But its role is better defined under pathological conditions such as Parkinson's disease. In patients with this condition, dopaminergic (dopamine-producing) neurons in the brainstem degenerate, resulting in a loss of dopamine throughout the basal ganglia. The dopamine loss in Parkinson's dis-

ease results in prominent oscillatory-burst discharges in the STN, GPi and SNr at oscillation frequencies of 3–8 Hz^{14,15}. Lesion studies¹⁶ indicate that these oscillations may contribute to the development of the tremor seen in patients with Parkinson's disease. Plenz and Kitai studied oscillatory-burst discharges in a co-culture environment containing neurons from the striatum, GP and STN, in which the different neuronal groups connect with each other appropriately. Because this co-culture did not contain dopaminergic cells, the culture conditions may better reflect the pathological condition (parkinsonism) rather than the normal state.

Neuronal oscillations as a property of local or extended networks are also thought to be important in some of the proposed rhythm generators within the thalamus and the cortex¹⁷. In these positive-feedback systems, reverberations may occur with resonant frequencies that depend mainly on transmission times and synaptic delays. But the STN/GPe system differs because the polarity of its connections makes it a negative-feedback system, which would dampen rather than facilitate oscillations (Fig. 1). This, and the fact that the observed primary oscillation frequencies are relatively low, indicates that oscillatory bursts in the STN/GPe network are probably not the result of a simple, reverberating circuit. Instead, these bursts seem to result from specific, pro-oscillatory membrane properties of STN neurons.

In support of this idea, Plenz and Kitai³, as well as previous studies⁸, show that oscillatory discharge can be induced in the STN neu-

rons by hyperpolarizing them with GABA (γ -aminobutyric acid). Many types of neuron in the basal ganglia share such membrane properties. Moreover, output nuclei of the basal ganglia participate in a variety of feedback loops, which may create the right conditions for oscillations to occur. So Plenz and Kitai's findings may be just the tip of the iceberg, with other potential independent oscillators in the basal ganglia still waiting to be discovered. Some may be involved in normal function, whereas others may contribute to pathological changes. □

Thomas Wichmann and Mahlon R. DeLong are in the Department of Neurology, Emory University, Suite 6000 Woodruff Memorial Research Building, 1639 Pierce Drive, Atlanta, Georgia 30322, USA. e-mail: twichma@emory.edu

1. Ivry, R. B. *Curr. Opin. Neurobiol.* **6**, 851–857 (1996).
2. Steriade, M. *Cerebral Cortex* **7**, 583–604 (1997).
3. Plenz, D. & Kitai, S. T. *Nature* **400**, 677–682 (1999).
4. Tiitinen, H. et al. *Nature* **364**, 59–60 (1993).
5. Chapman, C. A. & Racine, R. J. *J. Neurophysiol.* **78**, 2602–2615 (1997).
6. Jefferys, J. G., Traub, R. D. & Whittington, M. A. *Trends Neurosci.* **19**, 202–208 (1996).
7. Hamada, I. & DeLong, M. R. *J. Neurophysiol.* **68**, 1859–1866 (1992).
8. Beurrier, C., Congar, P., Bioulac, B. & Hammond, C. *J. Neurosci.* **19**, 599–609 (1999).
9. Nambu, A. & Llinas, R. *J. Neurophysiol.* **72**, 1127–1139 (1994).
10. Plenz, D. & Kitai, S. T. *J. Neurophysiol.* **76**, 4180–4184 (1996).
11. Aosaki, T., Kimura, M. & Graybiel, A. M. *J. Neurophysiol.* **73**, 1234–1252 (1995).
12. Raz, A., Feingold, A., Zelanskaya, V., Vaadia, E. & Bergman, H. *J. Neurophysiol.* **76**, 2083–2088 (1996).
13. Li, Y. X., Bertram, R. & Rinzel, J. *Neuroscience* **71**, 397–410 (1996).
14. Bergman, H., Wichmann, T., Karmon, B. & DeLong, M. R. *J. Neurophysiol.* **72**, 507–520 (1994).
15. Wichmann, T. et al. *Exp. Brain Res.* **125**, 397–409 (1999).
16. Bergman, H., Wichmann, T. & DeLong, M. R. *Science* **249**, 1436–1438 (1990).
17. Rowat, P. F. & Selverston, A. I. *J. Comp. Neurosci.* **4**, 103–127 (1997).

Ecology

Sprucing up beaver meadows

Peter D. Moore

In ecology there is no such thing as community spirit. Cooperation and altruism, we are told, are illusory, or simply the outcome of the activity of selfish genes. Yet the interdependence of species in communities is often extremely complex. One example is the complexity that characterizes the ecology of beaver ponds once they have been abandoned by the animals that first constructed them. This is discussed in the journal *Oikos* by John Terwilliger and John Pastor¹, who set out to discover why it takes so long for old beaver ponds to revert to forest.

When beavers dam streams and create their own aquatic environment, they inadvertently set in motion a directional process that eventually leads to re-establishment of the forest. The ponds begin to accumulate silt and are abandoned by the beavers. Then, once the dams are breached, the ponds drain and are soon invaded by herbaceous vegetation, especially grasses and sedges. This is

how the beavers create grassy openings in the forest — in much the way that high winds and fire create clearings, which are subsequently recolonized by species that prefer open habitats. But with fire and tree falls, healing is usually rapid, resulting in the invasion and growth of a new tree generation within a few years. Old beaver ponds, on the other hand, may remain as damp grassland patches for decades — often 70 years or so — before trees that can tolerate these wet conditions, such as black spruce (*Picea mariana*), eventually manage to establish a forest cover.

Why should this vegetation succession become stuck at the grassland stage, and not progress smoothly in its development? In most successions, 'facilitation' is observed. This is where certain species, simply by existing at a site, modify the local environment and make it more suitable for invasion by other species. Plants with associated nitrogen-fixing bacteria, for example, enrich the

soil with nitrogen², preparing the way for the invasion of more competitive and demanding trees. Could the beaver meadow be deficient in some nutrient? Or is a particular facilitator missing? Some successions stop at a grassland stage because of complex competitive interactions involving soil microbes. In savannah grasslands, for example³, some grasses may secrete chemical compounds that inhibit nitrification (the microbial conversion of ammonium ions to nitrate in the soil), thereby holding back the supply of nitrogen to potential competitors. Does such a microbially mediated process operate in the beaver clearings?

Terwilliger and Pastor¹ have examined this possibility, but concentrate their attention on mycorrhizal fungi. The black spruce is an ectomycorrhizal species — that is, it requires specific fungi in its root system for survival and growth. Clearings in the spruce forest created by wind or fire would retain the required fungi (or their viable spores) in the soil, allowing spruce seedlings to recolonize immediately. But beaver meadows have been submerged under water for some time. Long-term submergence and oxygen starvation are likely to be fatal to the fungi, so reinvasion of the spruce may be limited by the time it takes for fresh inoculum to arrive. Indeed, Terwilliger and Pastor confirm that sterilized spruce seeds sown and germinated in beaver-meadow soils failed to become infected with mycorrhizal fungi. Although the seedlings could survive without mycorrhizal fungi, growth was considerably poorer than that of fungus-infected seedlings planted in control soils from surrounding forests.

But the beaver meadows do eventually become invaded by spruce, so how are they inoculated with the required fungi? If it were simply a case of aerial spore dispersal, the fungi should arrive early in the succession. The long delay suggests that the fungal movements are very slow. Terwilliger and Pastor believe that such slow and chancy dispersal could occur if the fungi were carried by a small mammal that feeds in the forest, where it picks up the fungus and periodically moves out into the grassland — namely the red-backed vole (*Clethrionomys gapperi*; Fig. 1). This animal is deterred from frequent excursions into the grass-dominated ecosystem because that habitat is occupied by the meadow vole (*Microtus pennsylvanicus*). Conversely, the meadow vole does not venture into the forest, so it cannot pick up the fungus.

The red-backed vole is often found in forest-edge habitats, and its faeces have been shown to contain the spores of mycorrhizal fungi derived from its foraging in the forest. Terwilliger and Pastor inoculated beaver-meadow soils with *C. gapperi* faeces, and found that the spruce-tree seedlings were able to grow and survive under these conditions. The authors obtained the faeces from voles caught in either May or August, and



Figure 1 The red-backed vole (*Clethrionomys gapperi*). Terwilliger and Pastor¹ have found that these voles may carry fungal spores in their faeces — spores that facilitate the recolonization of beaver meadows by trees such as the black spruce.

found that the August-derived faeces were richer in spores and enhanced the growth of young spruce seedlings better than those obtained in May. Trapping records showed that red-backed voles were more likely to wander into the meadow during August.

Other factors could be involved in the slow succession of beaver meadows. One is the relatively high water table (sometimes within 10 cm of the soil surface). But these conditions should not be too moist for growth of the black spruce. Moreover, Ter-

williger and Pastor show that the ectomycorrhizal fungi could survive in the fluctuating water table of this hydrological regime — so, the soil wetness alone cannot account for the absence of the required fungi and the blocked succession. Grazing by meadow voles could eliminate the spruce seedlings. But then, why should this occur in beaver meadows and not in other grass-dominated clearings where meadow voles are found?

Whatever the answers, Terwilliger and Pastor's findings support the hypothesis that the red-backed vole is vital in facilitating the invasion of spruce trees into beaver meadows. The complexity of the interrelationships here, involving mammals, microbes and trees, provides a fine example of interdependence of organisms within a community, but also one in which chance, in the form of the wanderings of a vole, determines the speed of succession. □

Peter D. Moore is in the Division of Life Sciences, King's College, Franklin-Wilkins Building, 150 Stamford Street, London SE1 8WA, UK. e-mail: peter.moore@kcl.ac.uk.

1. Terwilliger, J. & Pastor, J. *Oikos* 85, 83–94 (1999).

2. Bormann, B. T. & Sidle, R. C. *J. Ecol.* 78, 561–578 (1990).

3. Rice, E. L. & Pancholy, S. K. *Am. J. Bot.* 59, 1033–1040 (1972).

Game theory

Agreeing on strategies

Ivar Ekeland

John F. Nash obtained one of the most basic results in game theory when he was in his early twenties¹. In 1950, he published a short paper on “Non-cooperative games”, which 45 years later was to earn him part of a Nobel Prize in Economic Sciences. Nash's equilibrium has permeated most of the social sciences and evolutionary biology, and is now standard fare in all textbooks on economic theory. It seems hard to improve on such a fundamental result, especially one based on such simple mathematics. Indeed, Nash's work was curtly dismissed by the great John von Neumann, the mathematician who, together with the economist Oskar Morgenstern, was the first to study strategic interaction between rational agents, and to coin the name game theory²: “This is trivial, you know. It's just a fixed-point theorem”. Even so, new work by Philip J. Reny³ from the University of Pittsburgh, published in *Econometrica*, unexpectedly improves on Nash's result.

First, let me describe the Nash equilibrium model. In a situation where a number of rational agents with different interests have to reach individual decisions by taking into account the decisions of others as well as their own, then the problem of correctly anticipating the actions of others becomes crucial. A Nash equilibrium is a situation where this has been achieved — that is, everyone turns out to have acted in their own

best interests in a way that correctly anticipates the actions of everyone else.

We would love to know a dynamic process whereby such a desirable situation could be reached. But we do not. Instead, Nash gave a set of conditions under which such an equilibrium exists, leaving aside the question of how to get there. Among these conditions there is a continuity requirement: the pay-off of every individual — the advantage she derives from any global outcome — must depend continuously on the various decisions taken. In other words, if she, or anyone else, slightly changes their actions, the new outcome, from her point of view (and from everyone else's), must be similar to the old one. Under this assumption, and another, more technical one (compactness and convexity of certain sets), a Nash equilibrium will exist.

Once reached, the equilibrium is stable, because no one has an incentive to deviate. I know that my last move was the best I could have made in response to how the others were playing, and I know that everyone else is in the same situation. So if we are asked to play one more time, everyone will be able to confidently repeat their previous move, and we will find ourselves in the same Nash equilibrium.

Unfortunately, the continuity requirement is not satisfied in many situations of real interest. For example, in a duel, firing just before or just after my opponent will make an

enormous difference to my chances of getting hurt. In a sealed-bid auction, when the bids are very close, small differences can change the winner. There is no way Nash's result can be twisted to accommodate such situations, however common they are, and there was no way to prove that an equilibrium existed until now. This is what Reny has done³. He turns his attention away from the pay-offs, and concentrates on the players' best responses to the others' actions. He defines a game as 'better-reply secure' if, whenever it is out of equilibrium, some player could change his action to get a better pay-off (up to now, this is simply stating that we are not in a Nash equilibrium), and that this new action would still guarantee him the higher pay-off, even if all the other players were to slightly change their own actions (this is Reny's new condition). He then shows that if a game is better-reply secure, and if it satisfies some technical conditions (compactness and convexity of certain sets, basically the same requirements as for Nash's old theorem), then an equilibrium exists.

Reny's theorem introduces Nash equilibria in situations where Nash's own result does not. Take two examples. First is a game of timing ('noisy duel'), where two opponents (of different skills) walk towards each other on a straight line, each having one shot (the duel is 'noisy' because both can hear the shot fired by the other). These games are used to model patent races, for example. Here Reny's condition is satisfied, and the game has a Nash equilibrium (which amounts to saying that there is a precise moment in time when each player should fire).

The second is an auction, where bidders compete for a single indivisible prize. Each bidder receives a signal, which is used to appraise the bounty. Different bidders receive

independent signals, and no one sees the signal except the recipient; but everyone knows the probability distribution of all signals, and how each signal will affect the actions of the recipient. The highest bidder wins the auction, pays his bid and walks away with the prize. Note that the simple-minded strategy of simply bidding your own appraisal may not be a Nash equilibrium, because you may find yourself overbidding everyone else: reducing your own bid would still yield the prize, while saving money. Reny's theorem applies to this situation, and shows that there is indeed a Nash equilibrium, which means that there is a way for all participants to exploit the public information to bid optimally.

Of course, the main problem with Nash equilibria is still there: they may exist, but how does one reach them? There are two answers. The first way is by brute calculation; which requires that the agents have enormous computational power at their disposal, and are confident that their opponents are doing the same. The second route is by trial and error, possibly extending over generations; there is no guarantee, however, even with extremely long-lived agents, that such a process will ever reach an equilibrium. We are in a situation akin to the beginning of mechanics: we can do the statics, but we don't have the dynamics. Even so, Reny's result stands out; it is a significant improvement on Nash's original work, and I am sure that it will prove as deep and fruitful as its predecessor. □

Ivar Ekeland is at the Institut de Finance, Université Paris-Dauphine, 75775 Paris, CEDEX 16 France.

e-mail: ivar.ekeland@dauphine.fr

1. Nasar, S. *A Beautiful Mind* (Faber & Faber, London, 1998).
2. von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behaviour* (Princeton Univ. Press, 1948).
3. Reny, P. J. *Econometrica* 67, 1029–1056 (1999).

Science in art

Turning the other cheek



Imagine you are going overseas and wish to have a portrait done as a gift for your family. You want them to remember you as

warm-hearted and affectionate. How would you pose?

This question was put to 165 psychology students by Michael Nicholls and colleagues, who report in *Proceedings of the Royal Society* (266, 1517–1522; 1999) that most of the students posed with their left cheek turned towards the camera. When asked to try and look unemotional and intelligent, however, the students were more likely to present their right cheek.

The authors found a similar leftward bias in portraits by a number of artists. They believe that, in informal portraits, sitters present their left cheek because this side of the face is controlled by the 'emotive' right cerebral hemisphere. But when Nicholls *et al.* looked at portraits of scientists, such as the one of Louis Pasteur shown here, they found the right cheek turned towards the artist in every single case.

Alison Mitchell

Daedalus

Eclipse of the Earth

Inspired by yesterday's eclipse of the Sun over Europe and Asia, Daedalus wants to set up a permanent eclipse. It could be sited at one of the Lagrangian points of gravitational stability in the Earth–Sun system, at which a spacecraft would 'keep station' relative to the Earth and Sun.

The outer Lagrangian point L2, 1,500,000 kilometres beyond the Earth along the Earth–Sun axis, seems the most promising choice. Seen from that position, the Earth subtends an angle of very nearly half a degree, as does the Sun. At L2, therefore, the Sun is almost perfectly obscured by the Earth. It would provide a permanent 'eclipse of the Earth'.

This novel eclipse would give splendid new insights into the Earth's atmosphere. Seen from L2, the Earth would be surrounded by a bright halo of sunlight, transmitted and scattered by its atmosphere. The light would be reddened, and would suffer selective absorption by dust and specific molecules in the air. In addition, it would suffer radial refraction. Rays which skim the Earth's surface are refracted round the Earth by almost a degree in the process. Those whose closest approach is high in the atmosphere are less refracted. So the Earth should show a radial rainbow or 'glory' of atmospheric light, encoding its density and composition as a function of height and latitude. As the Earth turned, the spectrum would sample the sunrise and sunset regions of the atmosphere over the whole globe. Within and behind this display, the night side of the Earth would emit its own less dazzling radiant signature. It would include the glow of lightning and aurorae, and infrared from natural and artificial thermal emissions.

The resulting spectra and images would be complicated and enriched by the Moon. A platform at L2 would keep station, not with the Earth, but with the centre of mass of the Earth–Moon system. The Earth itself librates monthly through this point. So much of the time the eclipse would be partial, with a sector of the Sun visible beyond one limb of the Earth, and the other showing a shifted pattern of atmospheric refraction. This useful monthly 'scan' would give meteorologists and atmospheric scientists even more detailed information. A TV camera relaying the whole spectacle back to Earth would also please the lay public, who have been taught that eclipses are dangerous to the health and should never be observed directly.

David Jones

Slaying the crystal homunculus

The existence, or otherwise, of atoms was hotly debated for several centuries. In the early 1800s, Eilhardt Mitscherlich's descriptions of crystal structures enraged some, but struck a decisive blow for the 'atomist' cause.

Robert W. Cahn

Do atoms exist? And what makes a crystal? Early in the nineteenth century, soon after John Dalton announced his atomic hypothesis, these long-running and contentious issues in physical science became dramatically linked. It is often forgotten that the study of certain types of crystal was one of the key verifications of the atomic hypothesis.

The great French crystallographer René-Just Haüy (1743–1822), for whom the taxonomy of mineral species was a life's work, believed that crystal forms were determined by *molécules intégrantes*, which he envisioned as tiny polyhedra. These were the smallest units into which a crystal could be broken down — the physical equivalent of the seventeenth-century view that a living embryo developed from a miniature 'homunculus' in the fertilized egg.

Haüy's crystal dogma overturned the much older theory of Kepler, Hooke and Huygens that one might explain the disposition of crystal faces in terms of the regular stacking of spherical atoms. Haüy convinced the world of mineralogy that crystals could not be understood in such terms, and that therefore there were no such entities as spherical atoms.

So it was not surprising that he attacked with sustained venom the astonishing findings of a German amateur mineralogist and professional chemist, Eilhardt Mitscherlich (1794–1863), who in 1818 found that the crystal forms of hydrated potassium phosphates and arsenates were identical. He went on to explore the same phenomenon among the sulphates and selenates. It is hard to appreciate today what a huge spanner this observation, of a phenomenon that came to be known as isomorphism, threw into the sedate machinery of nineteenth-century chemistry and mineralogy.

The young Mitscherlich took a long time to find his vocation. He began by studying oriental languages, before moving on to medicine and from there to chemistry, inspired by a Berlin chemistry professor, Heinrich Link, who was renowned as a man of wide interests. Both men were driven to find connections between apparently disparate themes, and Link urged his pupil on to the study that led to the recognition of isomorphism. To make a good job of it, Mitscherlich had to learn the techniques of mineralogy, and turned to an expert, Gustav Rose, for instruction.

Subsequently, Mitscherlich met the great Swedish chemist Jöns Berzelius (1779–1848), who took him under his wing and persuaded



Mitscherlich: his discovery of isomorphism caused a stir in the sedate world of mineralogy.

the German authorities to appoint him, at a tender age, to a chair in Berlin. Mitscherlich's research on phosphates and arsenates was published more fully in 1821 in Swedish, and a German translation became a classic. There is an exact reprint of this translation, dated 1898, in the series *Ostwald's Klassiker der Exakten Wissenschaften*. (Wilhelm Ostwald was the father of physical chemistry, born a few years before Mitscherlich died.)

In the same 1821 paper, Mitscherlich also recognized the existence of polymorphs of the same substance, by reference to the classic puzzle of calcite and aragonite, both of which have the composition CaCO_3 . Later, he studied polymorphism and dimorphism — the existence of quite different crystal forms of the same substance — in sulphur. Haüy had twisted himself into knots trying to show that the two calcium carbonate polymorphs were not chemically identical, because polymorphism was incompatible with his model of the *molécules intégrantes*.

A third finding, made almost simultaneously with the discovery of isomorphism, by François Beudant in France and William Wollaston in England, was that isomorphous species can form a series of solid solutions with each other, what German miner-

alogists came to name *Mischkristalle*, mixed crystals. This was also incompatible with the existence of *molécules intégrantes*.

Mitscherlich and Berzelius, respectful of Haüy as a great experimental scientist, tried for years to persuade him of the validity of their findings, but he was implacable. Berzelius finally declared that "one ought not to expect that a grey-haired scientist close to the end of an honourable life should give up a theory he erroneously considered to be the most important of his discoveries; this is perhaps too much to morally demand of any man".

Berzelius declared Mitscherlich's discovery and interpretation of isomorphism, jointly with Pierre-Louis Dulong and Alexis-Thérèse Petit's discovery in France that the specific heats of solids vary inversely with their (presumed) atomic weights, as the most important empirical proofs of the atomic hypothesis at that time. Yet there was widespread scepticism about atoms for another century, until Jean Perrin's work on Brownian motion early in the twentieth century finally routed the remaining disbelievers. □

Robert W. Cahn is in the Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK.

DEUTSCHES MUSEUM, MUNICH

Convergent evolution of cytokines

Functional analogues of vertebrate inflammatory cytokines have been described in a variety of invertebrates¹. The analogy is based mainly on the crossreactivity of antibodies elicited against vertebrate cytokines, the sensitivity of invertebrate immunocytes to the action of vertebrate cytokines, and the responsiveness of vertebrate immune cells to invertebrate factors. But without knowing the amino-acid or gene sequences of the putative invertebrate cytokine analogues, it has not been possible to demonstrate unequivocally a phylogenetic relationship between vertebrate cytokines and their invertebrate functional analogues. Here we show that, although a defence molecule from the earthworm *Eisenia foetida* (Oligochaeta, Annelida) and the mammalian tumour-necrosis factor TNF- α perform similar functions, they emerged independently during evolution.

Although cytokines exert their activity by interacting with specific receptors, there is some evidence that they participate in innate immune functions through lectin-like interactions². In particular, TNF- α has been shown to contain a lectin-like domain (the TIP domain) that is spatially distinct from its TNF-receptor-binding site and which mediates the lytic activity of TNF- α on salivarian trypanosomes³. Furthermore, the binding of TNF- α to microbial saccharides may provide an opsonin-like signal that mediates the attachment of bacteria to macrophages². The lectin-like activity of TNF- α may therefore be of physiological importance during the vertebrate defence reaction to microorganisms.

To test for invertebrate molecule(s) with lectin specificity similar to TNF- α , we used murine monoclonal antibodies elicited against the TIP domain⁴ to detect putative immunoreactive component(s) in the coelomic fluid of *E. foetida*. The anti-TIP monoclonal antibody reacts with a protein of relative molecular mass 42,000 that has been described as the glucan- and lipopolysaccharide-binding protein CCF-1 (for coelomic cytolytic factor-1), which is

Table 1 Inhibition of trypanolytic activity of CCF-1 and TNF- α

	Trypanolysis (%) mediated by:	
	TNF- α	CCF-1
No inhibitor	51 $\pm$ 5	56 $\pm$ 4
<i>N,N'</i> -diacetylchitobiose	0	4 $\pm$ 1
Cellobiose	47 $\pm$ 3	50 $\pm$ 6
Anti-TIP	0	2 $\pm$ 0.5
Anti-CCF-1	1 $\pm$ 0.4	3 $\pm$ 0.4
Control antibody	45 $\pm$ 5	58 $\pm$ 6

Inhibitors (2 μ g ml⁻¹) were preincubated for 1 hour at 25 °C with concentrations of TNF- α or CCF-1 (purified from coelomic fluid by immunoaffinity) that induced about 50% lysis (10⁴ pg ml⁻¹ TNF- α or 5 $\times$ 10³ pg ml⁻¹ CCF-1).

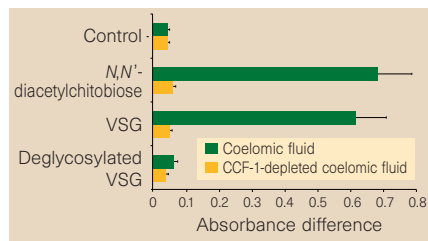


Figure 1 Prophenoloxidase activation in *Eisenia foetida* coelomic fluid by *Trypanosoma brucei* components. Coelomic fluid was incubated with 0.1 μ g ml⁻¹ *N,N'*-diacetylchitobiose, variant-specific glycoprotein (VSG) or VSG treated with *N*-glycosidase F. The level of the prophenoloxidase activation was estimated as the absorbance difference (mean $\pm$ s.d.) of 3,3,4-dihydroxyphenyl-L-alanine oxidation⁵.

involved in activating the invertebrate prophenoloxidase defence mechanism^{5,6}. CCF-1 can be isolated from *E. foetida* coelomic fluid on immobilized *N,N'*-diacetylchitobiose, a saccharide specific for the lectin-like domain of TNF- α , but not on immobilized cellobiose, for which TNF- α has no affinity. Furthermore, a monoclonal antibody elicited against CCF-1 reacts with TNF- α . CCF-1 (GenBank accession number AF030028) and TNF- α (Swissprot accession number P06804) do not share any sequence similarity (analysed using Genetics programs).

TNF- α lyses *Trypanosoma brucei* through its TIP domain^{3,4} and CCF-1 has a domain and lectin activity similar to that of TNF- α . We tested the coelomic fluid of *E. foetida* for trypanolytic activity and found that it lysed bloodstream forms of *T. brucei*. Depletion of CCF-1 from *E. foetida* coelomic fluid on an anti-CCF-1 monoclonal antibody immunoaffinity column abolishes this lytic activity, whereas immunoaffinity-purified or recombinant CCF-1 lyse *T. brucei*. The trypanolytic activity of CCF-1 and TNF- α is impaired by *N,N'*-diacetylchitobiose, anti-TIP and anti-CCF-1 monoclonal antibodies (Table 1), indicating that CCF-1, like TNF- α , interacts with African trypanosomes through a *N,N'*-diacetylchitobiose lectin-like domain, causing lysis of the parasites.

The biological significance of the trypanolytic activity of CCF-1 remains unclear, although recognition of *N,N'*-diacetylchitobiose by CCF-1 may be important for invertebrate defence mechanisms mediated through the activation of prophenoloxidase^{5,6}. Indeed, *N,N'*-diacetylchitobiose, a component of Gram-positive bacteria and yeast cell walls, initiates the prophenoloxidase cascade in whole, but not in CCF-1-depleted, coelomic fluid (Fig. 1). These data support the idea that CCF-1

acts as a recognition molecule by triggering the prophenoloxidase defence mechanism in *E. foetida*⁵.

T. cruzi epimastigotes are reported to trigger the prophenoloxidase cascade in crayfish⁷. The variant-specific glycoprotein (VSG) that acts as a protective coat on bloodstream forms of *T. brucei* has an *N*-linked *N,N'*-diacetylchitobiose core and induces activation of the prophenoloxidase cascade in *E. foetida* coelomic fluid (Fig. 1). CCF-1 is involved in the prophenoloxidase activation by VSG because CCF-1 depletion abolishes the oxidative activity of the coelomic fluid triggered by VSG. This activity is restored by the addition of recombinant CCF-1. Digestion of VSG with *N*-glycosidase F fully impairs its capacity to trigger prophenoloxidase activation (Fig. 1), indicating that CCF-1 interacts with the *N,N'*-diacetylchitobiose core of VSG to induce the phenoloxidase cascade in *E. foetida* coelomic fluid. Finally, preincubating VSG with TNF- α impairs the activation of prophenoloxidase in *E. foetida* coelomic fluid. Taken together, these results show that CCF-1 and TNF- α use a similar *N,N'*-diacetylchitobiose-lectin-like activity to perform distinct yet inherent biological functions: prophenoloxidase activation in invertebrates, and trypanolytic activity in mammals.

We have previously reported that CCF-1 is secreted by *E. foetida* coelomocytes upon stimulation with lipopolysaccharide⁸, whereas in mammals lipopolysaccharide triggers the production of TNF- α by macrophages. Both CCF-1 and TNF- α have opsonizing properties^{2,9} and bind β -1,3 glucan through a lectin-like interaction^{5,10}. We have shown that CCF-1 and TNF- α share an *N,N'*-diacetylchitobiose lectin-like activity that may have been functionally conserved as a recognition mechanism during innate defence reactions in invertebrates and vertebrates. However, despite their functional analogies, CCF-1 and TNF- α do not show genetic homology, indicating that they lack a common evolutionary origin. CCF-1 cannot therefore be considered as an invertebrate cytokine homologue.

Alain Beschin*, Martin Bilejt, Lea Brys*, Els Torreele*, Rudolf Lucas*, Stefan Magez*, Patrick De Baetselier*

*Department of Immunology, Parasitology and Ultrastructure, Flemish Interuniversity Institute for Biotechnology, VIB-VUB, Paardenstraat 65, 1640 St-Genesius-Rode, Belgium
e-mail: abeschin@vub.ac.be

†Department of Immunology, Institute of Microbiology, Academy of Sciences of the Czech Republic, Videnska 1083, 142 20 Prague 4, Czech Republic

- Habicht, G. S. & Beck, G. *Adv. Comp. Environ. Physiol.* **24**, 29–47 (1996).
- Lucas, R., Magez, S., De Leys, R. & De Baetselier, P. in *Lectins: Biology, Biochemistry, Clinical Biochemistry* (eds Van Driessche, E., Fischer, J., Beeckmans, S. & Bog-Hansen, T.C.) 244–249 (Textop, Hellerup, Denmark, 1994).
- Lucas, R. *et al. Science* **263**, 814–817 (1994).
- Magez, S. *et al. J. Cell Biol.* **137**, 715–727 (1997).
- Beschin, A. *et al. J. Biol. Chem.* **273**, 24948–24954 (1998).
- Smith, V. J. *Adv. Comp. Environ. Physiol.* **23**, 75–114 (1996).
- Barracco, M. A. & Söderhall, K. *Braz. J. Med. Biol. Res.* **29**, 1321–1327 (1996).
- Bilej, M. *et al. Immunol. Lett.* **60**, 23–29 (1998).
- Bilej, M. *et al. Immunol. Lett.* **45**, 123–128 (1995).
- Olson, E. J., Standing, J. E., Griego-Harper, N., Hoffman, O. A. & Limper, A. H. *Infect. Immun.* **64**, 3548–3554 (1996).

Rise in carbon dioxide changes soil structure

Carbon in soil affects the formation and stabilization of aggregates (groups of primary particles that adhere to each other more strongly than to surrounding soil particles)¹. Soil aggregation is important for preventing soil loss through wind and water erosion, and the size distribution and abundance of water-stable aggregates influences a range of physical, chemical, biological and agricultural properties of soil². The effects on soil biota and nutrient cycling of increases in soil carbon availability, brought about by increased CO₂, are well studied, but the consequences for soil aggregation and structure have not been examined. Here we show for three ecosystems that the water stability and size distribution of aggregates is affected by long-term CO₂ fumigation, and we propose a mechanism for this that involves the production by fungi of the glycoprotein glomalin.

The Jasper Ridge CO₂ experiment in northern California³ exposed two natural annual grassland ecosystems (sandstone and serpentine) to increased atmospheric CO₂ for six growing seasons by using cylindrical, open-top chambers (1 m tall, 0.33 m², *n* = 10). In both grasslands, a higher proportion of soil was found in aggregates 1–2 mm across in elevated CO₂, and the proportion of aggregates of 0.25–1 mm was significantly increased in the sandstone

grassland (Table 1). The water stability of both size classes followed a pattern similar to the mass of aggregates. This suggested that the higher mass of aggregates could be explained by an increase in the water stability of aggregates (Table 1).

Although soil aggregation is a complex hierarchical process⁴, the soil concentration of the glycoprotein glomalin⁵ is tightly correlated with aggregate stability across many soils⁶. Glomalin is produced mainly by hyphae of arbuscular mycorrhizal fungi⁷, which form symbiotic associations with plant roots. The length of the hyphae in these fungi increases with elevated CO₂ in the sandstone grassland, but not in the serpentine grassland, with root biomass and length showing the opposite pattern⁷. Total glomalin and immunoreactive glomalin concentrations in soil increased in both grasslands with elevated CO₂ (Table 1). Glomalin concentration in aggregates (from a separate extraction) increased under elevated CO₂ for aggregates of 0.25–1 mm in both communities, but this was not the case for those of 1–2 mm (Table 1). The water stability of that fraction may be under different control.

The Sky Oaks CO₂ study in southern California used 12 greenhouses (2 × 2 × 2 m) with controlled CO₂, ambient lighting and controlled temperature at six CO₂ concentrations from a pre-industrial level of 250 μl l⁻¹ to 750 μl l⁻¹ at intervals of 100 μl l⁻¹ (*n* = 2). The chambers were built around *Adenostoma fasciculatum* (chamise) shrubs in chaparral vegetation recovering from an experimental burn. Soil samples were taken after three years of treatment and analysed for soil aggregation and glomalin concentration to see whether the patterns in the grasslands also existed in a different vegetation type. The proportion of soil mass in aggregates of 0.25–1 mm showed a linear increase (linear regression, *P* = 0.03, *r*² = 0.74) along the CO₂ gradient, but the 1–2 mm aggregate mass did not (*P* = 0.68, *r*² = 0.04). Glomalin concentrations followed a pattern similar to that of the small aggregate size class (*P* = 0.03, *r*² = 0.71).

The carbon sink represented by glomalin over the experimental period for Jasper

Ridge was 8.29 g C m⁻² in the serpentine and 4.25 g C m⁻² in the sandstone grassland. These are very small amounts compared with the large organic carbon stocks in these soils, and are on the order of 5% of the total calculated litter and soil accumulation under elevated CO₂ on an annual basis⁸. Glomalin therefore seems to be more important in carbon sequestration by virtue of its function in soil aggregation (which has been linked with carbon stabilization) than by acting as a carbon sink itself.

Our results indicate that changes in soil structure in response to CO₂ enrichment should be incorporated into global research because soil structure has a strong effect on soil processes and organisms. On a global scale, the extent of soil degradation and erosion is severe⁹ and is accelerated by changes in many global factors, including climate and land use¹⁰. Our finding that an increase in soil aggregation could be brought about by atmospheric change may have implications for studies of soil stabilization in ecosystems.

Matthias C. Rillig*, Sara F. Wright†, Michael F. Allen‡, Christopher B. Field*

*Department of Plant Biology, Carnegie Institution of Washington, Stanford, California 94305, USA
e-mail: matthias@jasper.stanford.edu

†US Department of Agriculture, Agricultural Research Service, Beltsville, Maryland 20705, USA

‡Center for Conservation Biology, University of California, Riverside, California 92521, USA

- Kemper, W. D. & Rosenau, R. C. in *Methods of Soil Analysis, Part I* (ed. Klute, A.) 425–442 (American Society of Agronomy, Madison, Wisconsin, 1986).
- Coleman, D. C. *Fundamentals of Soil Ecology* (Academic, San Diego, 1996).
- Field, C. B., Chapin, F. S., Chiariello, N. R., Holland, E. A. & Mooney, H. A. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 121–145 (Academic, San Diego, 1996).
- Tisdall, J. M. & Oades, J. M. *J. Soil Sci.* **33**, 141–163 (1982).
- Wright, S. F. & Upadhyaya, A. *Soil Sci.* **161**, 575–586 (1996).
- Wright, S. F. & Upadhyaya, A. *Plant Soil* **198**, 97–107 (1998).
- Rillig, M. C., Allen, M. F. & Field, C. B. *Oecologia* **119**, 572–577 (1999).
- Hungate, B. A. *et al. Nature* **388**, 576–579 (1997).
- Daily, G. C. *Science* **269**, 350–354 (1995).
- Valentin, C. in *Global Change and Terrestrial Ecosystems* (eds Walker, B. & Steffen, W.) 317–338 (Cambridge Univ. Press, 1996).

Table 1 Effects of increased CO₂ on aggregates and glomalin in two annual grasslands

	Sandstone		Serpentine		P-values (ANOVA)	
	Ambient CO ₂	Increased CO ₂	Ambient CO ₂	Increased CO ₂	CO ₂	Grassland
Aggregates 1–2 mm (% of soil)	14.43 (0.54)	15.14 (0.33)	16.70 (0.56)	18.08 (0.50)	0.04	0.50
Aggregates 0.25–1 mm (% of soil)	17.06 (0.58)	20.04 (0.82)	26.93 (1.62)	25.61 (1.09)	0.46	0.05
Water stable 1–2 mm (%)	86.70 (1.58)	90.10 (1.27)	76.15 (1.47)	81.02 (1.26)	0.002	0.06
Water stable 0.25–1 mm (%)	88.87 (0.90)	92.77 (0.63)	84.27 (0.85)	85.32 (0.49)	0.006	0.87
Total glomalin (mg g ⁻¹)	2.09 (0.04)	2.17 (0.03)	2.60 (0.05)	2.79 (0.06)	0.01	0.26
Immunoreactive glomalin (mg g ⁻¹)	0.67 (0.04)	0.75 (0.02)	0.97 (0.02)	1.06 (0.04)	0.01	0.97
Glomalin, 1–2 mm (mg g ⁻¹ ag)	1.83 (0.04)	1.85 (0.05)	2.40 (0.05)	2.36 (0.14)	0.93	0.71
Glomalin, 0.25–1 mm (mg g ⁻¹ ag)	2.08 (0.04)	2.32 (0.05)	2.50 (0.09)	2.62 (0.09)	0.02	0.34

Values in brackets are standard error of the mean (*n* = 10). *P* values (obtained by analysis of variance (ANOVA); suitable transformations were used as necessary) of less than 0.05 are in bold. Water stability of aggregates (expressed as the percentage of stable aggregates) was measured using a wet-sieving method following capillary rewetting of soil samples. Glomalin was extracted by repeated autoclaving in a citrate extraction buffer. Glomalin concentrations were measured by immunoreactivity assay using an enzyme-linked immunosorbent assay with monoclonal antibody 32B11 against glomalin isolated from arbuscular mycorrhizal fungal hyphae. Total glomalin concentration in extracts was measured using a Bradford protein assay. Glomalin concentrations are expressed per gram of soil or aggregate (ag). Further details are available from the authors.

Anisotropy of iron in the Earth's inner core

Seismological studies^{1–3} suggest that there is 3–4% anisotropy of P-waves in the Earth's inner core. Hexagonal closed-packed solid iron has been proposed to be the major constituent of the inner core⁴, but a lack of knowledge about the elastic properties of this phase at inner-core pressures (330 to 360 GPa) and temperatures (4,000 to 8,000 K) prevents a conclusive interpretation of the seismic results in terms of the inner-core composition. The elastic properties of hexagonal closed-packed iron have been computed from first principles using a theoretical method⁴ and Mao *et al.*⁵ have recently measured them by using X-ray diffraction and ultrasound techniques at pressures of up to 211 GPa. Despite these exciting technological achievements, we believe that the new results may be misleading.

Although the theoretical and experimental results both indicate similar aggregate P-wave and S-wave velocities as a function of pressure⁵, there is an inconsistency in the elastic constants that Mao *et al.*⁵ have attempted to resolve by multiplying the P-wave velocities at 16 and 211 GPa by 1.08. However, this treatment results in a 15–40% change in elastic parameters at 211 GPa and is accompanied by a 7–30% decrease in the S-wave velocity (Fig. 1). The

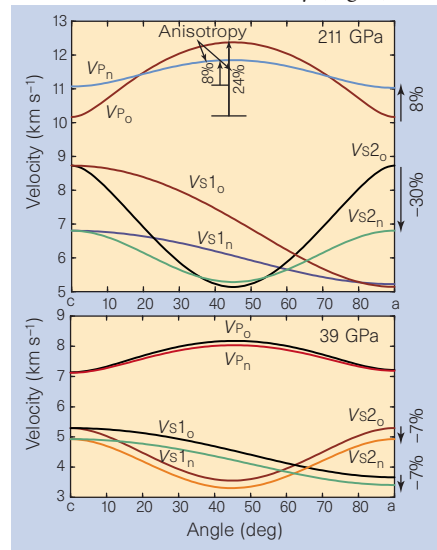


Figure 1 P-wave and S-wave velocities (V_p and V_s , respectively) as a function of the propagation direction relative to the c -axis of the hexagonal close-packed iron crystal for the original and corrected tables of Mao *et al.*⁵. The subscript 'o' denotes values derived from the original table and 'n' denotes those from their amended table. The percentage increase or decrease in velocity from the original values is indicated. The difference between the two results is significant: the P-wave anisotropy has decreased from 24% to 8%.

authors assume that the values at 39 GPa are correct, but even at that pressure some of the elastic parameters have changed by up to 15% (Fig. 1). Furthermore, in their correction, Mao *et al.*⁵ obtain a value for the elastic constant C_{33} of 491 GPa, which is 63% less than the value of 802 GPa obtained at 39 GPa pressure⁴. Such a low value for C_{33} yields a very low Poisson's ratio (0.04) along the symmetry axes. Mao *et al.*'s⁵ value for C_{44} is 38% higher than that obtained at 211 GPa pressure⁴.

Our analysis suggests that the uncertainty in the individual elastic constants is much greater than in parameters based on a combination of them, such as the P-wave velocity. Given the large uncertainties and the physical implications of the elastic parameters, the results of Mao *et al.*⁵ must be interpreted with caution.

S. C. Singh*†, J.-P. Montagner*

**Département de Sismologie, Institut de Physique du Globe de Paris, 4 Place Jussieu, 75252 Paris, France*

†*Bullard Laboratories, University of Cambridge, Cambridge CB3 0EZ, UK*

1. Morelli, A., Dzwiewonski, A.M. & Woodhouse, J. H. *Geophys. Res. Lett.* **13**, 1545–1548 (1986).
2. Woodhouse, J. H., Giardini, D. & Li, D. *Geophys. Res. Lett.* **13**, 1549–1552 (1986).
3. Shearer, P. M., Toy, K. M. & Orcutt, J. A. *Nature* **333**, 228–232 (1988).
4. Stixrude, L. & Cohen, R. E. *Science* **267**, 1972–1975 (1996).
5. Mao, H. *et al. Nature* **396**, 741–743 (1998); correction, *ibid* **399**, 280 (1999).

Nitrogen deposition and carbon sequestration

Nadelhoffer *et al.*¹ use ¹⁵N-tracer studies in nine northern forests to argue that increasing inputs of combined nitrogen from the atmosphere are unlikely to cause the increase in forest growth that has been postulated as the 'missing sink' for atmospheric CO₂. Only about 20% of the tracer ended up in the trees and about 70% remained in the organic and mineral layers of the soil. If only 20% of the nitrogen input from the atmosphere were available for tree growth, then not enough combined nitrogen would be coming into the northern forests each year to explain the missing carbon sink.

We do not agree with this argument for several reasons. First, the ¹⁵N-tracer was applied to the soil surface in all the experiments, minimizing the opportunity for assimilation by soil organisms living in the nitrogen-poor organic litter layer. But much of the combined nitrogen received by trees from the atmosphere will be taken up directly by leaves, without ever reaching the ground, irrespective of whether it arrives in rain, as an aerosol or in gaseous form^{2,3}. The efficiency of use of the combined nitrogen

received by forests will therefore be greater than the 20% used by Nadelhoffer *et al.* in their calculations.

A second problem arises in the interpretation of experiments using ¹⁵N-labelled fertilizers. In unfertilized soil, assimilation of nitrogen by microorganisms that decompose nitrogen-poor plant debris is met from the unlabelled inorganic nitrogen pool. In fertilized soil, the same amount of assimilation occurs (assuming that assimilation is driven by microbial demand rather than the size of the pool), but some of this assimilated nitrogen is labelled. This means there is more unlabelled nitrogen and less labelled nitrogen left in the pool for the trees to take up, a process known as pool substitution. The net effect is that the percentage of ¹⁵N taken up by the trees underestimates the true uptake. This underestimate is more serious⁴ when the pool of unlabelled mineral nitrogen is small and when the soil contains much decomposing organic matter, as in most of the experiments of Nadelhoffer *et al.* A decline in the proportion of ¹⁵N recovered by plants as fertilizer application decreases is an indication⁵ that pool substitution has occurred (see their Fig. 1).

Nadelhoffer *et al.* may be right in saying that not enough combined nitrogen reaches the northern forests to explain the missing sink, but their analysis of ¹⁵N-labelled-fertilizer data is not proof. They claim that a maximum of 20% of the combined nitrogen reaching the forest is absorbed by trees, but we think this is a minimum value.

Finally, Nadelhoffer *et al.* imply that retention of carbon in tree biomass leads to longer-term sequestration than retention in soil. But this depends on the fate of the trees, as part of the carbon from tree litter entering the soil is likely to be held in soil carbon fractions that have turnover times of several centuries⁶.

D. S. Jenkinson, K. Goulding, D. S. Powlson

Soil Science Department, IACR-Rothamsted, Harpenden, Hertfordshire AL5 2JQ, UK
e-mail: keith.goulding@bbsrc.ac.uk

1. Nadelhoffer, K. J. *et al. Nature* **398**, 145–148 (1999).
2. Asman, W. A. H., Sutton, M. A. & Schjørring, J. K. *New Phytol.* **139**, 27–48 (1998).
3. Stulen, I., Perez-Soba, M., de Kok, L. J. & van der Eerden, L. *New Phytol.* **139**, 61–70 (1998).
4. Jenkinson, D. S., Fox, R. H. & Rayner, J. H. *J. Soil Sci.* **36**, 425–444 (1985).
5. Hart, P. B. S., Rayner, J. H. & Jenkinson, D. S. *J. Soil Sci.* **37**, 389–403 (1986).
6. Smith, P. *et al. Geoderma* **81**, 153–225 (1997).

From their ¹⁵N tracer results¹ and an estimated deposition rate to forests of 5.1×10^{12} g nitrogen per year^{2,3}, Nadelhoffer *et al.* conclude that about 0.25×10^{15} g carbon per year is sequestered by northern temperate forests. Their ¹⁵N-tracer measurements are robust but there is a problem with their assumption that they have simulated atmospheric nitrogen deposition at forests by adding ¹⁵N tracers systematically to forest

floors "in all cases". Atmospheric nitrogen deposition is actually intercepted by the forest canopy, particularly at the nine closed-canopy forests studied.

It is known that forest canopies chemically interact with HNO_3 , NO_x and NH_3 gases (among others), as well as with the NO_3^- and NH_4^+ ions in particulate matter. The Integrated Forest Study⁴, covering 13 sites with a wide range of North American and European forest types and atmospheric nitrogen deposition, varying from 5 to 30 kg per hectare per year, reported that "an average 40% of the incoming inorganic N is retained in or transformed at passage through the canopy; the net canopy exchange is greatest at those sites receiving the highest atmospheric N input." At some sites in this study in the eastern United States, the canopy retention of atmospheric inorganic nitrogen deposition is very high: for example, at Howland Forest, Maine⁵, nitrate canopy retention is about 90% and ammonium canopy retention is more than 80%. Retained nitrogen deposition can also be rapidly assimilated by canopy foliage⁶.

The ^{15}N applied by Nadelhoffer *et al.*¹ to the forest floor would have bypassed the canopy retention and assimilation pathway. The nitrogen deposition allocated to the non-woody and woody biomass ecosystem pools (15 and 5%, respectively; Table 2 of ref. 1) may have been significantly increased if it had been possible to apply the labelled nitrogen to the forest canopy.

Doubling or tripling the percentage of nitrogen deposition allocated to woody biomass (10 or 15%, rather than the 5% used) would increase forest carbon uptake by 50–100% over that estimated by Nadelhoffer *et al.* The woody-biomass nitrogen allocation may be increased several-fold, and could be determined by using labelled nitrogen applied to the forest canopy.

Using a figure of 5.1×10^{12} g per year as the global nitrogen deposition to forests may also lead to an underestimate. It is recognized that there is much uncertainty in the estimates of forest nitrogen deposition^{2,3}, particularly in the case of $\text{NH}_x\text{-N}$ deposition, which may be substantially greater^{7,8} than the 2×10^{12} g nitrogen per year considered in ref. 1. Total anthropogenic nitrogen deposition has been estimated to be as much as 18×10^{12} g per year to temperate and boreal forests⁹, which is three or more times the value reported in ref. 1. These two points can provide an upper bound for estimates of forest carbon sequestration of $1\text{--}2 \times 10^{15}$ g carbon per year, with the range dependent on the still poorly known nitrogen deposition allocated to woody biomass.

Herman Sievering

Department of Environmental Sciences, Geography and Physics, Global Change and Environmental Quality Program, University of Colorado at Denver,

Denver, Colorado 80217, USA
e-mail: hsieveri@carbon.cudenver.edu

1. Nadelhoffer, K. J. *et al.* *Nature* **398**, 145–148 (1999).
2. Townsend, A. R., Braswell, B. H., Holland, E. A. & Penner, J. E. *Ecol. Appl.* **6**, 806–814 (1996).
3. Holland, E. A. *et al.* *J. Geophys. Res.* **102**, 15849–15866 (1997).
4. Johnson, D. W. & Lindberg, S. E. (eds) *Atmospheric Deposition and Forest Nutrient Cycling* (Springer, New York, 1992).
5. McLaughlin, J. W., Fernandez, I. J. & Richards, K. J. *J. Environ. Qual.* **25**, 248–259 (1996).
6. Calanni, J. *et al.* *Environ. Pollut.* **105**, 79–89 (1999).
7. Schlesinger, E. H. & Hartley, A. E. *Biogeochemistry* **15**, 191–211 (1992).
8. Galloway, J. N., Levy, H. & Schlesinger, W. H. *Glob. Biogeochem. Cycles* **9**, 235–252 (1995).
9. Hudson, R., Gherini, S. & Goldstein, R. A. *Glob. Biogeochem. Cycles* **8**, 307–333 (1994).

Nadelhoffer *et al.* reply — Jenkinson *et al.* and Sievering are justifiably concerned that our ^{15}N additions to forest floors do not account for the potential uptake of nitrogen input by forest canopies. We agree that canopies can remove nitrogen from the atmosphere, resulting in inputs to forest floors that are less than the total nitrogen deposition. A North American study¹ has suggested that canopies remove, on average, 16% of total (organic + inorganic) atmospheric nitrogen input to forests, and concluded that nitrogen uptake by the canopy is probably small relative to the nitrogen requirements of trees. Spraying ^{15}N -labelled ammonium and nitrate on the crowns of five-year-old Norway spruce indicated that foliar uptake in mature forests probably constitutes only a small percentage of annual nitrogen uptake².

Although canopies can remove nitrogen from bulk deposition, the extent to which this nitrogen is biologically available and stimulates wood growth is poorly understood. Throughfall nitrogen fluxes at our sites either exceed the measured bulk nitrogen deposition or closely agree with modelled total deposition^{3–5}, so their use as a proxy for nitrogen deposition is reasonable and does not compromise our conclusions.

We think it is unlikely that "pool substitution" of unlabelled inorganic nitrogen for ^{15}N tracers in soils biased our estimates of throughfall uptake by trees, as claimed by Jenkinson *et al.*, who found this mechanism operating in theoretical models⁶ and in potted, but not in field-grown, plants⁷. Comparison of the amounts of our tracer additions to the amount of unlabelled inorganic nitrogen already present in our soils with those amounts in their theoretical analysis (see Fig. 1 of ref. 6) indicates that pool substitution is unimportant at the low ^{15}N enrichment used in our throughfall manipulations.

We agree that forest soils can serve as long-term carbon sinks and that carbon turnover times in soil pools are longer than in trees. Nevertheless, as nearly all carbon enters forests through trees, it is here that the influence of nitrogen inputs on carbon uptake is important. We point out, however, that our simple stoichiometric budget

(Table 2 of ref. 8) actually factors in the role of soils, with soil storage calculated as the product of the C:N ratio and nitrogen immobilization in soils.

Finally, we recognize that there is a wide range of estimates for nitrogen deposition on temperate forests. This is because of uncertainties and likely but unknown biases in the atmospheric mixing models, emissions data and site-specific wet + dry nitrogen deposition data used to generate spatial predictions of nitrogen deposition on forests. Our temperate-forest deposition value of 5.1×10^{12} g nitrogen was derived from a comparison of modelled NO_x and NH_x deposition estimates⁹. If this value were to be in serious disagreement with estimates based on greatly improved monitoring and modelling efforts, then our assessment of the effects of nitrogen deposition on temperate-forest carbon uptake would need to be revised.

Future estimates of this effect must also consider spatially explicit patterns of nitrogen loss from forests in relation to nitrogen deposition and land use¹⁰, the likelihood of narrowing C:N ratios under increased nitrogen input, and possible nutritional imbalance of tree tissue subjected to high nitrogen input. Meanwhile, our tracer experiments indicate that, although nitrogen deposition in forests accounts for some of the northern-temperate CO_2 sink, other factors must account for most of this sink.

Knute J. Nadelhoffer*, **Bridget A. Emmett†**,
Per Gundersen‡, **Chris J. Koopmans§**,
Patrick Schleppi||, **Albert Tietema¶**,
Richard F. Wright#

*The Marine Biological Laboratory,
Woods Hole, Massachusetts 02543, USA
e-mail: knute@mbi.edu

†Institute of Terrestrial Ecology,
Deiniol Road, Bangor LL57 2UP, UK

‡Danish Forest and Landscape Research Institute,
Hoersholm Kongevej 11, 2970 Hoersholm, Denmark
§Louis Bolk Institute, Hoodstraat 24,
3972 LA Driebergen, The Netherlands

||Swiss Federal Institute for Forest, Snow and
Landscape Research, WSL, Zürcherstrasse 111,
8903 Birmensdorf, Switzerland

¶University of Amsterdam, Nieuwe Prinsengracht 130,
1018 VZ Amsterdam, The Netherlands

#Norwegian Institute for Water Research,
Box 173 Kjelsaas, 0411 Oslo, Norway

1. Lovett, G. M. in *Atmospheric Deposition and Forest Nutrient Cycling* (eds Johnson, D. W. & Lindberg, S. E.) 159–166 (Springer, New York, 1992).
2. Wilson, E. J. & Tiley, C. *Atmos. Environ.* **32**, 513–518 (1998).
3. Currie, W. S., Aber, J. D., McDowell, W. H., Boone, R. D. & Magill, A. H. *Biogeochemistry* **35**, 471–505 (1996).
4. Schleppi, P. *et al.* *Water Air Soil Pollut.* (in the press).
5. Tietema, A., Emmett, B. A., Gundersen, P., Kjønaas, O. J. & Koopmans, C. *Forest Ecol. Mgmt* **101**, 19–27 (1998).
6. Jenkinson, D. S., Fox, R. H. & Rayner, J. H. *J. Soil Sci.* **36**, 425–444 (1985).
7. Hart, P. B. S., Rayner, J. H. & Jenkinson, D. S. *J. Soil Sci.* **37**, 389–403 (1986).
8. Nadelhoffer, K. J. *et al.* *Nature* **398**, 145–148 (1999).
9. Holland, E. A. *et al.* *J. Geophys. Res.* **102**, 15849–15866 (1997).
10. Aber, J. D. *et al.* *BioScience* **48**, 921–934 (1998).

Culture clash in American anthropology

Culture: The Anthropologists' Account

by Adam Kuper
Harvard University Press: 1999. 299 pp.
\$29.95, £18.50

Mary Douglas

Satire is hard to bear; the more elegant it is, the more it hurts. A quarter of a century ago, Adam Kuper wrote a critical history of British social anthropology. As he admits in the preface to the third edition of *Anthropology and Anthropologists: The Modern British School* (Routledge), he was young and heedless then, and he was astonished by the fury that greeted the book in England. But the Americans loved it. It is only fair that Kuper should now write a critical history of recent American anthropology, a hatchet job if ever there was one. What may not be exactly fair is that this second history of the subject, *Culture: The Anthropologists' Account*, is a mature work, profound, well researched, and therefore even more lethal than the first.

English 'social anthropologists' between the 1920s and 1970s had broken away from the current perception of their discipline, turning their backs on culture and mythology to become absorbed in the systematic comparison of social organization. In complete contrast, the American anthropologists, every bit as seriously, claimed 'culture' to be their special province. Re-establishing the discipline after the Second World War, the professors kept trying to define it, but the task proved difficult because, by itself, the word is empty. In the eighteenth- and nineteenth-century philosophical debates in France, Germany and England, the word 'culture' figured prominently, always steeped in political bias and always changing its tinge. With post-structuralist hindsight, it is obvious that, in its old usage, culture was part of a contrast set, sometimes the good part, as in a contrast of civilized/primitive or spiritual/material, sometimes the bad part, as in a contrast of artificial/simple. When the American anthropologists wanted a neutral concept for the focus of an objective intellectual discipline, they innovated by making culture stand on its own. But, true to its history, the term remained apt for polemic. As American anthropology developed, some preferred the cultural approach to interpretation simply because it was untainted by economic determinism.

In America, culture — that is, ideas and values — came to be treated as a separate field, the life of the mind, independent of economics, politics and psychology. As the book unfolds, this trend towards bracketing culture away gets stronger, leading implicitly to cultural determinism and so setting the



A difference of perspective: the circumstances surrounding the death of Captain Cook in Hawaii in 1779 remain a subject of anthropological controversy.

terms for debate between Marxist and cultural explanations. Eventually, the patterns of ideas and values that are the object of anthropological research come to be divorced from the patterns of work and action that interest the social scientists. Without particularly intending it, or even noticing, the discipline left the ground and floated up into philosophical idealism. One result is to impede dialogue between idealist anthropology and the realist social sciences. Both academic spheres are the poorer for the lack of interaction. For their part, the British anthropologists still remain strongly entrenched in a realist mode by their involvement with social studies. No wonder conversation is awkward between British and American colleagues.

From the UK side of the Atlantic, the recent American controversy over the death of Captain Cook in Hawaii seems totally mysterious. The question is the extent to which behaviour is governed by mythic formulas or by rational-instrumental decisions. With such defective records about a single event 200 years ago, how could one ever hope to settle it? The idealist perspective actually hampers the articulation of these problems.

To illustrate his general thesis, Kuper closely examines the work of three leaders of the present generation of anthropologists: two giants, Clifford Geertz and Marshall Sahlins, and one stormy petrel, the late David Schneider. The first two come out with honours, both having conducted exemplary and fascinating enquiries into economics and politics as well as kinship and religion, Geertz in Islamic countries, Indonesia and Morocco, Sahlins in Polynesia. Geertz only becomes vulnerable when he seeks to give his views on political realities in the countries he studied. Sahlins is more seriously vulnerable because the philosophical problems of representation are among his central interests, so the idealist starting point is a more fundamental drawback. It would seem that initial

vaguenesses about professional research objectives and about the concept of culture do not prepare the anthropologist well for policy analysis or philosophy.

Cool irony inspires Kuper's keen scrutiny, and anthropologists not mentioned can only feel grateful for their exclusion. More significantly, as the last part of the book shows, the profession is now strongly radicalized in the sense that tough theoretical issues have become marginalized, while good causes are brought centre stage.

Anthropology is becoming an expressive branch of social work, attending to human rights, the defence of oppressed persons and environments, the post-Colonial aftermath and the politics of identity. Kuper argues throughout that the dispersion of scholarly effort can be blamed on the "resolutely idealist conception of culture". One can add that this standpoint creates intellectual tangles that discredit the subject, call into question the importance of the academic enterprise and weaken the resistance of professionals to pressure from politics and the media.

Steven Shapin's superb *Social History of Truth* (University of Chicago Press, 1994) shows that, in the seventeenth century, the trustworthiness of a scientist's theory partly depended on his social and religious credentials. Now we witness the secular version: the acceptability of an American anthropologist's work partly depends on his political credentials.

On a professional note, Kuper has betrayed an idealist's bias in his treatment of the subject. To provide a social rather than a cultural anthropology, he would have had to have analysed the institutional underpinnings of the profession. All the same, it remains an important story. The author need not worry this time around about the anthropologists' reaction. Even the targeted victims will welcome it, and not because Americans enjoy seeing their fellow countrymen brought low, nor because they are

generous to critics. The distinguished achievements in the history of ideas will be acclaimed, although the story of the self-destruction of a once-respected, learned discipline is saddening. □

Mary Douglas is at 22 Hillway, Highgate, London N6 6QA, UK.

What did you do in the [cold] war, Daddy?

Biohazard

by Ken Alibek

Random House: 1999. 319 pp. \$24.95, £17.99

Malcolm R. Dando

In his well-known history, *The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present* (Harper-Collins), Roy Porter argued: "... the latter part of the nineteenth century brought one of medicines's few true revolutions: *bacteriology* [emphasis added]. Seemingly resolving age-old controversies over pathogenesis, a new and immensely powerful aetiological doctrine rapidly established itself ...". Porter went on to point out that, very unusually for medicine, this revolution brought dramatic and rapid benefits to the human population in new preventative measures and remedies. Unfortunately, the demonstration — by scientists of the standing of Louis Pasteur and Robert Koch — that specific diseases were caused by specific microorganisms, also raised the possibility that the new knowledge might be misused in offensive biological-warfare programmes. We know now that, during the First World War, both sides attempted to use biological weapons to sabotage the other side's valuable animal stocks.

Subsequently, during the middle of the twentieth century, other advances in biology and medicine — such as in aerobiology and production microbiology — were used in major offensive biological-warfare programmes by countries such as the United Kingdom and United States. For some years it appeared that such misuse of science had been halted by agreement of the Biological and Toxin Weapons Convention (BTWC) in the early 1970s. But the convention lacked any effective verification conditions, and it has recently become clear that the former Soviet Union embarked on a vast expansion of its biological-weapons programme at the very time the convention was agreed.

A proper description of the full extent of this modern programme is not available in the public domain. From a variety of sources, however, we know some of its characteristics. First, we know that it was carried out on a massive scale, with numerous institutes and many thousands of people involved. Second, it involved the large-scale creation of weapons involving a range of agents, and so

implied a willingness to use biological weapons in major military operations. Finally, it is clear that recent advances in genetic engineering were being used — for example to increase the antibiotic resistance of plague.

Ken Alibek was one of the leading scientists and organizers of this offensive biological-weapons programme and, through a series of interviews on television and in newspapers and magazines, he has been a principal source of public knowledge about what was done. Alibek's book, *Biohazard*, ghost written by Stephen Handelman, gives a readable account of what can reasonably be called a 'chilling true story' in the form of an autobiography.

I feel strongly that this book should be compulsory reading for everyone involved in the dramatic biotechnology/genomics revolution today. Many other scientific advances have been applied in new weapons systems, and we will be fortunate indeed if modern biology is not misused in the same way.

Although Alibek's account may be fallible in parts, enough has been confirmed by other sources for the whole to be taken very seriously. Different readers will be struck by different parts of this story. For me, the account of project 'Bonfire' was particularly alarming. Alibek describes being in a long, boring review meeting in 1989. One of the last speakers was due to report on this project, which was a long-running attempt to genetically engineer a human pathogen to produce an additional toxin or bioregulator. Alibek recounts: "The test was a success. A single genetically engineered agent had produced symptoms of two different diseases, one of which could not be traced. The room was absolutely silent. We all recognized the implications ... A new class of weapons had been found ..."

He goes on to describe how such new weaponry might be used to damage heart function or to target the nervous system and behaviour. Today's neuroscientists, striving to find means of helping those afflicted by mental problems, might also be concerned by the related 'Flute' programme, which was devoted to developing psychotropic agents to induce mood and behavioural changes in people for malign purposes.

Nevertheless, Alibek comes over in the book as a thoughtful and decent man. We could all perhaps learn from his experience. In his final paragraph he states: "... I cannot unmake the weapons I manufactured or undo the research I authorized as scientific chief of the Soviet Union's biological weapons programme ..."

He continues: "... but every day I do what I can to mitigate their effects. The realization that even today, in Iraq or China, another father of three may be sitting down at a conference table to plot the murder of millions of people is what spurs me on ..."

We should all ask ourselves whether we

have done enough to help secure the agreement of the BTWC Verification Protocol now being negotiated in Geneva, for this is surely the best way available to us of preventing such misuse of biology and medicine. □

Malcolm R. Dando is in the Department of Peace Studies, University of Bradford, Bradford BD7 1DP, UK.

Neural networks beyond Freud

The Mind Within the Net

by Manfred Spitzer

MIT Press: 1999. 359 pp. \$27.50, £19.50

Terrence J. Sejnowski

Sigmund Freud was a pioneer in a field that today is called neural networks. Before turning to the 'talking cure', he sought explanations for normal as well as disordered thought processes in the flow of 'nervous energy' through networks of neurons, based on what was known about the neurophysiology and anatomy of the brain at the end of the nineteenth century. From these early speculative ideas came such concepts as repression and sublimation which would become the foundation of psychoanalysis. In *The Mind Within the Net*, psychiatrist Manfred Spitzer takes us back to these roots and asks whether recent advances in cognitive neuroscience and neural-network models provide a firmer foundation for explaining the mysteries of human experience.

Many of the important technical breakthroughs that fuelled the neural-network revolution that began in the 1980s were made by psychologists, as heralded in the two volumes on *Parallel Distributed Processing*, edited by David Rumelhart and James McClelland (MIT Press, 1986). A new mathematical foundation was developed for explaining human cognition based on models with continuous variables and dynamical systems rather than computer programs based on logic and symbol processing. Spitzer does not cover these latest developments, but has written a highly readable introduction to 'traditional' neural-network models.

Thinking about how populations of neurons encode the sensory world and motor coordination, and how they acquired these properties, is not easy. Computer simulations of relatively simple neural-network models have revealed their powerful ability to solve complex problems, such as the recognition of facial expressions. The same networks that were built to mimic normal human behaviour can also be used to explore the inexplicable breakdown patterns that occur when brains are damaged. *The Mind Within the Net* excels at exploring a wide range of clinical problems, including phantom limbs, autism and schizophrenia.



Wiring up: neural-network models have thrown light on how we perceive and experience our world.

To the extent that the models can be related to brain systems, they can be tested and, eventually, the chasm that once separated the mind and the brain can be bridged.

One of the most surprising and puzzling recent discoveries in brain science has been the degree to which the cerebral cortex, the most highly evolved part of the human brain, remains malleable. The surface of the body is mapped onto the surface of the cortex in such a way that nearby points on the body map onto nearby neurons in the cortex. When a sensory nerve that innervates the fingers of a monkey is cut, the map reorganizes and the cortical area once dedicated to that patch of the body surface is, over time, reassigned to neighbouring body surfaces. This process involves both cortical and sub-cortical mechanisms for neural plasticity. Conversely, when a monkey is asked to use its fingers repeatedly over a long period, the area devoted to those fingers enlarges. Something similar happens in the somatosensory cortex of human Braille readers.

Neural-network models of cortical maps with Hebbian synaptic plasticity on the sensory inputs can reproduce the changes that occur during loss of neuronal input. Shortly after limb amputation in humans, vivid phantom sensations can occur, often accompanied by intractable pain that is referred to the missing limb. Curiously, reports of phantom limbs are rare following spinal injury leading to paraplegia. Why should these two ways of cutting off neuronal input to the cortex have such different perceptual consequences?

One possibility, explored by Spitzer in a cortical model, assumes that there is noise in the stump of the severed nerve. In the model, the noise excites cortical neurons and leads

to cortical reorganization but, in the case of spinal injury where it is assumed that such noise is absent or reduced, the reorganization of the map in the model is attenuated. The advantage of the model is that plausible hypotheses can be generated and explored in a system whose complexity escapes direct reasoning.

Reading this book was like visiting an old friend after a long absence. The friend is ageing gracefully but has taken up new hobbies. Since these simple neural-network models were introduced we have learned much more about the brain, and a new generation of neural-network models has been explored, based on the detailed biophysical properties of neurons. Despite this greater sophistication, many of the concepts that emerged from the first generation of neural-network models are still applicable, such as attractor states and population codes. One major difference is a richer dynamic by virtue of the intrinsic properties of the neurons. With much faster computers it has also been possible to simulate more neurons and to test how the dynamical states scale with the size of the network.

Perhaps the most puzzling property that these new network models exhibit, mimicking the brain, is spontaneously active oscillations. New concepts are needed to explain the complex mixtures of rhythms that occur in these networks and to characterize their computational power. Perhaps ten years from now in a second edition Spitzer or his students will draw out the consequences of these new models for psychology and psychiatry in an equally clear and effective way.

The last chapter contains a brief "User's manual for your brain". Just as cardiologists can provide advice on diet and exercise to enhance your physical well-being, Spitzer

provides a mental diet and exercise that may enhance brain development in children and maintain a sharp mental edge in adults. It should be possible to expand this user's manual as further discoveries are made about the brain and integrated into a clearer picture of how brains are influenced by the world.

Freud provided a language for organizing human experience that lasted nearly 100 years. A new language of the mind is now emerging from cognitive neuroscience and computer models that may have far-reaching consequences for how we see ourselves. □

Terrence J. Sejnowski is at the Salk Institute and in the Department of Biology, University of California, San Diego, California 92093, USA.

The first three billion years

Cradle of Life: The Discovery of Earth's Earliest Fossils

by J. William Schopf
Princeton University Press: 1999. 367 pp.
\$29.95, £17.95

Stefan Bengtson

Of all the fossils on Earth, some have to be the earliest. So what's the big deal, except for an entry in *The Guinness Book of Records*? Quite a big deal, in fact. The Phanerozoic — the 'time of visible [animal] life' of the past 550 million years — is now known to have been preceded by three billion years or more of Precambrian life, almost exclusively microbial. It seems that life appeared as soon as conditions at all permitted it. Life may well be (to use Christian de Duve's words) a cosmic imperative.

Bill Schopf has been a prime mover on this frontier, in his quest to find microfossils by slicing up uncountable chunks of the Earth's older crust and in his messianic efforts to bring together scientists of all creeds and talents to ask: 'What does it all mean?' His persistent questioning, arguing, pleading, shouting, bullying, persuading, fund-raising, entertaining, writing and editing have probably promoted and inspired more interdisciplinary work on the history of life than any other factor.

In the well-written *Cradle of Life*, Schopf tells his story of how Earth's early microbial biosphere was discovered. He ranges from biochemistry to natural history, science history and personal anecdotes, and although the path can be tortuous, it is not too convoluted. The many diagrams, however, usually lack information about data sources; this is particularly troublesome, as it is apparent that some plots are based on idealized data.

Schopf's treatment of the earlier players in the field is balanced and entertaining, but currently active players hardly get a mention. One could argue with some of the judge-

ments. A. C. Seward is (dis)credited with stifling research in the field for almost 40 years through his 1931 critique of the known or asserted Precambrian fossil record. If Seward's call for "more critical examination of the evidence" indeed had such an effect, surely the blame lies with those who might have taken up the challenge but didn't.

Nor does Schopf spare himself; the account of how, in 1965, he helped his professor take advantage of a peer-review commission to get the edge on a competitor would raise more than eyebrows if the act were committed today. The 1983 paper by Awramik, Schopf and Walter, reporting the world's oldest fossils from the Warrawoona Group in western Australia, is called a fiasco, because the original sampling spot could not be found; Schopf's own 1993 paper is credited with the real discovery. However, the earlier work showed that these fossils existed, and so was no more a fiasco than Christopher Columbus's voyage (about which there was also uncertainty over exact locations). In this vein, Schopf's paper rather compares to Amerigo Vespucci's voyages. Isn't that good enough?

The book is not free from bloopers, which could have been avoided by more careful reviewing. The map of North America on page 37 should be deeply offensive to those of a Canadian persuasion in New Brunswick and Nova Scotia; the Bolsheviks are credited with seizing power from the czar (who had actually abdicated more than half a year before the Bolshevik coup); and atmospheric carbon dioxide is said to act "like the windows of a greenhouse, it holds in heat, storing it in the chemical bonds that knit its atoms together" (no, it absorbs infrared radiation and converts it to heat in the atmosphere; greenhouse windows mostly just keep warm air in).

A central message of the book is that Precambrian biology differs fundamentally from Phanerozoic biology. This is indeed so; the transition between the two stands out as the greatest revolution in the history of the biosphere. Schopf makes a claim for 'hypo-bradytely' — extremely slow evolution during the Precambrian, owing to the predominance of asexual reproduction, "a new fundamental insight that ... stands out as one of the most striking findings ... since Darwin". Whether or not this bold claim will stand the test of time, the measurement of evolutionary rates in fossil microorganisms faces formidable difficulties, because in general only morphology is preserved, and poorly so.

Another obstacle to understanding early fossils is the bias imposed by the small subset of life-forms that have happened to survive until today. By using living organisms as our models (for example, by identifying the earliest fossils as cyanobacteria, as Schopf does), we may blind ourselves to the host of extinct organisms, and unwittingly introduce hypo-

bradytely as a taxonomic artefact.

This problem follows us into astrobiology; we need methods to recognize life-forms, fossil or living, that may be totally unrelated to those on Earth. In an epilogue on that theme, Schopf mischievously places a chapter on two of the most famous howlers in palaeontology — Scheuchzer's 'Homo diluvii testis' and Beringer's 'Lügensteine' — side-by-side with a critical discussion and effective debunking of the claim for ancient life in the Martian meteorite ALH84001 (Schopf was the invited sceptic at the NASA press conference in 1996). In no way does he deride the field, however. Like Seward 65 years earlier, Schopf has advocated constructive scepticism in a field of research where even the existence of the subject is in doubt. Like Seward, he deserves praise for this. □

Stefan Bengtson is in the Department of Palaeozoology, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden.

Damsels and dragons

Dragonflies: Behaviour and Ecology of Odonata

by Philip S. Corbet

Harley: 1999. 882 pp. £62.50

Mike Siva-Jothy

Dragonflies have almost certainly been using the same habitats in the same way for more than 250 million years. But, despite the general conservatism in this insect order, there is considerable variation in many aspects of their larval and adult behaviour, much of which is influenced by ecology. Philip Corbet's book is a rare beast: a truly authoritative natural history monograph dealing with a large taxon, and not strictly the behavioural ecology text that the title suggests.

Almost 20 years ago, as a student, I heard a rumour that Corbet, the patriarch of odonatologists, was about to write a new book. As the years passed, expectations grew, but later turned into concern that perhaps the work would not be finished. Key contributors to the material and impetus for the book passed away while it was *in utero*, including the extraordinary Peter Miller, whose research contributes something to every chapter. Not surprisingly, when the book eventually appeared, there was concern that it would not live up to 18 years of anticipation. But it did. Every aspect of biology touched upon is explained clearly and carefully by an author who has truly made the effort to understand it all.

Corbet has synthesized a phenomenal amount of material. He deliberately does not organize the book around the currently important questions in evolutionary ecology. Instead he has written a cleverly cross-referenced and indexed text that provides all the



raw material from which ideas can be developed, or key information sought. The value in this approach is that, as current questions are replaced, the book will still provide well-organized information.

The book follows the sequence of the key life-history stages, starting with eggs, moving through larval behaviour and life-history traits, then dealing with emergence, pre-adult development and, in the largest chapter, reproduction. This is the most complete synthesis of behavioural, ecological and background material from a single taxonomic group I have ever seen. As though it were not enough to synthesize so much information, Corbet has also provided a large set of appendices giving valuable follow-up material from each chapter.

Corbet never pulls his linguistic punches, and uses strictly correct entomological terms throughout. This is initially daunting (as is the uncompromising use of binomial nomenclature), but the reasoning behind it is important: these terms are increasingly falling into disuse as well as misuse. The glossary at the end is invaluable.

I have only two quibbles. First, most of the figures are directly reproduced from their original sources, which means sometimes there is a hotchpotch of illustrative styles that does not always do justice to the text. Second, an illustrated key of the major taxa would have been a useful foil to the abundance of binomial names. But these criticisms are pebbles, thrown at an edifice that will sit next to Bert Hölldobler and Edward O. Wilson's *The Ants* (Springer, 1990) as one of the definitive natural history texts of the twentieth century. □

Mike Siva-Jothy is in the Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.



correction In the review "A user's guide to two minutes of totality" by Jay M. Pasachoff (*Nature* **399**, 651–652; 1999), the photograph of the total eclipse of the Sun was taken by the Extreme-ultraviolet Imaging Telescope Team, Solar and Heliospheric Observatory, NASA and ESA.

Streaks of microearthquakes along creeping faults

Allan M. Rubin*, Dominique Gillard*† & Jean-Luc Got‡

* Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA

‡ Laboratoire de Géophysique Interne et Tectonophysique, UMR CNRS, C5559 et Université de Savoie, 73376 Le Bourget-du-Lac, France

Crustal faults that produce most of their slip aseismically typically generate large numbers of small earthquakes. These events have generally been interpreted as coming from localized patches of the fault that undergo unstable (stick-slip) sliding, surrounded by larger regions of stable sliding (creep). In published catalogues the microearthquakes often appear to be distributed over large portions of the fault surface. By accurately locating large numbers of microearthquakes from faults of different orientations in California and Hawaii, we show here that instead the locations define highly concentrated streaks that are characteristically aligned in the direction of fault slip. The underlying cause of this structural organization of the fault surface remains to be determined.

Faults near the Earth's surface slip in various ways: in large earthquakes that rupture the entire 10–20 km thickness of the brittle crust, in microearthquakes orders of magnitude smaller than this thickness, in 'creep events' or 'slow earthquakes' that last for hours or days and so do not radiate seismic energy, and at the slow, steady velocities of tectonic plates¹. To an extent that is incompletely understood these behaviours are not mutually exclusive; for example, many faults that creep largely aseismically simultaneously produce abundant microearthquakes ('aseismic' in this sense means that the total slip rate represented by the earthquakes, when averaged over the fault surface, amounts to only a small fraction of the observed long-term slip rate). Because earthquake size follows a power-law distribution, with roughly 10 times as many earthquakes produced for each unit drop in magnitude (ref. 2), most of our information about the geometry of active faults comes from microearthquakes near the detection threshold of the local seismic network. Even in well instrumented areas, the location errors of such earthquakes are typically several hundreds of metres or more. This severely limits our ability to use microearthquakes to probe fault-zone structure; for comparison, the earthquake-generating cores of extinct fault zones exposed by mining or erosion are inferred to be less than tens of metres thick^{3,4}.

By obtaining much more accurate locations for large numbers of microearthquakes, it may be possible to image the internal structure of fault zones and see if this structure correlates with the very different styles of fault behaviour that are observed. Several methods for relocating microearthquakes based on a comparison of seismic waveforms have been developed in recent years^{5–9}. All take advantage of the fact that two earthquakes that occur near the same spot and that have similar focal mechanisms (that is, fault orientations and slip directions) will produce waveforms at a given seismic station that appear similar. By cross-correlating these waveforms, it is possible to obtain very accurate estimates of the relative arrival times, and hence the relative locations, of the two events.

Data analysis

We concentrate on the earthquakes located along the 30-km section of the San Andreas fault indicated in Fig. 1, stretching from the southern edge of the 1989 Loma Prieta rupture in the north to 11 km southeast of San Juan Bautista in the south. Our data set consists of all earthquakes recorded by the US Geological Survey Northern California Seismic Network (NCSN) between May 1984 and December 1997 with waveforms that may be obtained from the Northern California Earthquake Data Center. This amounts to

more than 3,200 events in the box shown. Only about 10% pre-date the Loma Prieta earthquake; most of these are located in the southern half of the box.

This region is of interest because it straddles the boundary between the creeping and the locked sections of the San Andreas fault near San Juan Bautista¹, and because it has been the site of several slow earthquakes recorded by borehole strainmeters since 1990 (refs 10–12). The creeping section, which extends south-eastwards to near Parkfield, slips largely aseismically at nearly the rate defined by the relative motions of the Pacific and North American plates (about 30 mm yr⁻¹). The locked section to the north slipped by about 2 m during the major 1906 earthquake¹³, but does not undergo measurable creep at the surface and typically does not produce large numbers of microearthquakes except as after-shocks of larger events. The recorded slow earthquakes occurred within the southern half of the box, and were comparable to magnitude 5 earthquakes in terms of their moment (proportional to the product of fault surface area and slip).

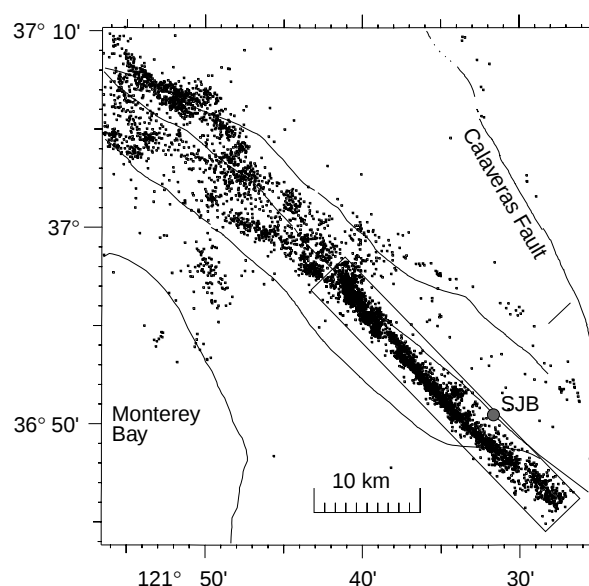


Figure 1 Map of the study area. The box encloses the San Andreas fault earthquakes examined in this study, occurring from May 1984 to December 1997. Events shown outside the box are limited to those occurring within 10 days of the 1989 Loma Prieta earthquake. SJB, San Juan Bautista.

† Present address: Exxon Production Research, PO Box ST-2158, Houston, Texas 77252, USA.

To relocate large data sets we have further automated the method described by Got *et al.*⁵. Starting with 1,000–1,500 earthquakes from the NCSN catalogue, we first identify families of similar earthquakes (multiplets) based on a comparison at each seismic station of the seismograms of all pairs of earthquakes located within 4 km of one another. The multiplet selection criteria are adopted by trial and error so as to enable us to relocate as large a fraction of the catalogue as possible while excluding events that are insufficiently similar for accurate relocation (see Methods section). For each multiplet we then measure, at each station, the difference in arrival time

(delay) between all pairs of sufficiently similar events. Measurements are made independently for compressional (P) and shear (S) waves. By performing the waveform cross-correlation in the frequency domain, delays can be measured with subsample (less than 0.01 s) accuracy. By comparing each event to many others, rather than to a single master event, we increase the accuracy of the relocation and are able to image laterally extensive structures. The internal consistency of the redundant delay measurements¹⁴ indicates that at the better stations we routinely obtain delays with accuracies of 0.1 to 0.2 samples (1 to 2 ms) across multiplets that

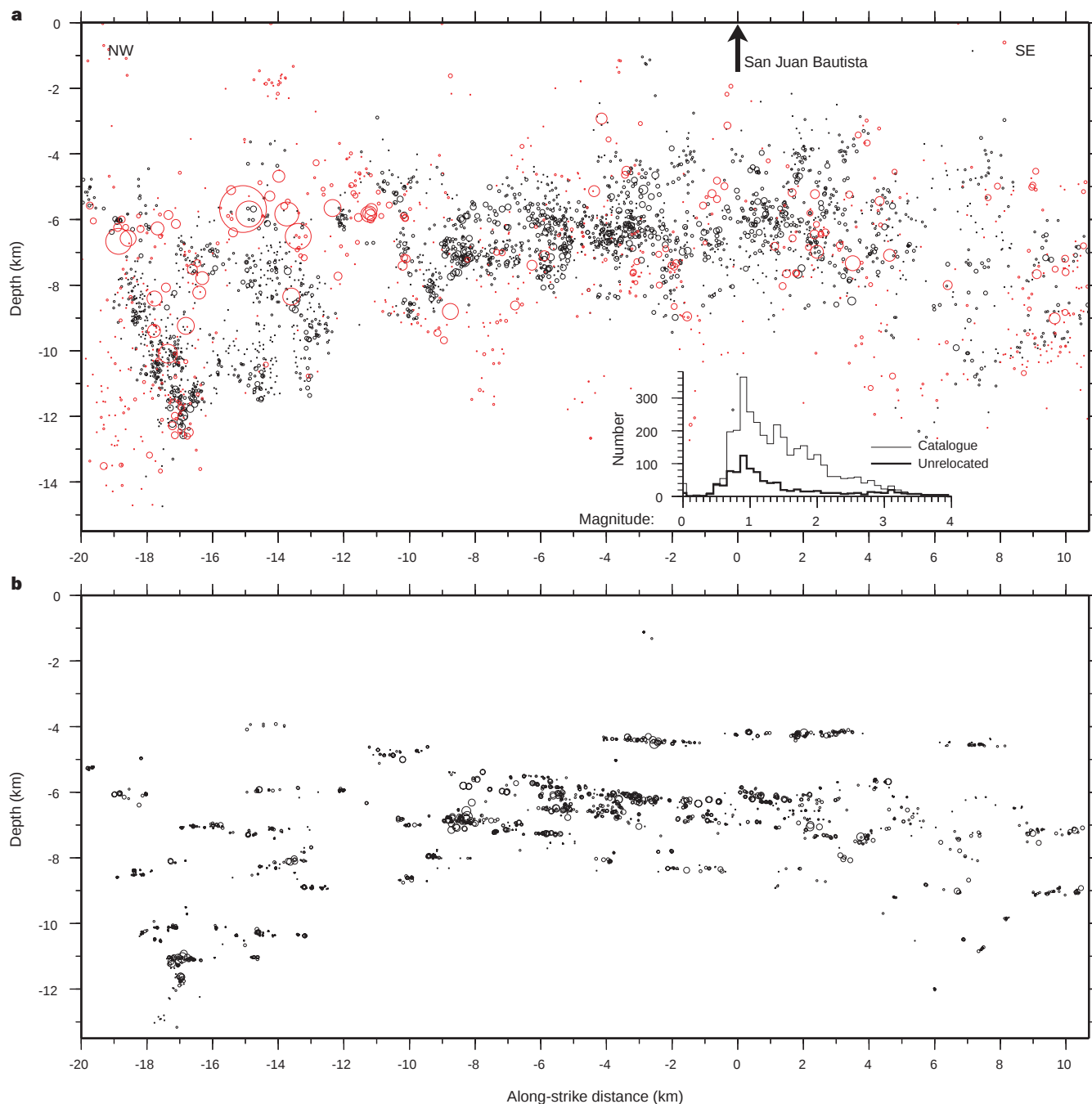


Figure 2 Vertical along-strike cross-sections showing the earthquakes examined in this study. Circles provide an estimate of event size, using the relation between moment (M_0) and magnitude (M); $\log(M_0) = 1.0M + 9.8$ (ref. 18). **a**, Locations from the NCSN catalogue of 2,410 earthquakes selected as belonging to multiplets (black) and 830 earthquakes not selected (red). Inset, the magnitude distribution of the entire catalogue (thin line) and those earthquakes not grouped into multi-

plets (thick line). **b**, The 65 multiplets following relocation. The absolute location of the centre of each multiplet is taken to be the average of the NCSN catalogue locations of its member events. The relocation shifted the earthquakes by median distances of 180 m horizontally and 300 m vertically, consistent with the relative magnitudes of horizontal and vertical location errors in the NCSN catalogue.

span hundreds of metres (ref. 15).

The final step is a linear inversion for the relative locations and origin times of the events that minimize the differences between the computed and observed delays. For these small earthquakes, the locations correspond to the event centroids. As described by Got *et al.*⁵, it is assumed that the events are sufficiently close together that the rays leaving all the events for a given station are parallel, and that differences in travel time may be ascribed solely to the different path lengths in the source region. The velocity model that is adopted influences the relocation only insofar as it determines the ray parameters and the velocity in the source region, with the source velocity affecting only the size and not the shape of the relocated multiplet. We have modified this method so that for multiplets larger than about 1 km, the ray paths are not treated as uniform across the multiplet but may vary according to the mid-point of each event pair under consideration. We also allow for different seismic velocities in the source region depending upon which side of the fault the station is on. It is well established that in this region the velocity of the Pacific plate is as much as 20% higher than that of the Franciscan rocks on the North American plate^{16,17}; this velocity difference is also apparent in our delay measurements.

The above procedure determines relative locations of earthquakes within a multiplet, but not the absolute location of that multiplet. This we take to be the average of the NCSN catalogue locations of its

member events. In a few cases, a single earthquake was placed into each of two large multiplets with centres separated by 1–3 km; the two ‘absolute’ locations obtained for these events generally coincide to within 200 m. This provides a reasonable estimate of the uncertainty in the relative position between adjacent multiplets.

Relocation results

We grouped more than 2,400 earthquakes, or 75% of our data set, into 65 multiplets of five or more events. In Fig. 2a we show the NCSN catalogue locations of both the multiplet (shown black) and non-multiplet (shown red) events. Note that in several kilometre-scale regions the multiplets contain nearly 100% of the catalogue seismicity. The circles provide an indication of earthquake size, based on a moment–magnitude relation for small earthquakes in California¹⁸. We assume circular ruptures and identical (3 MPa) stress drops such that an event of magnitude 1.5 has a radius of 30 m.

Following relocation the multiplets define a remarkably linear, nearly horizontal fabric (Fig. 2b). One lineation (at horizontal position $x = -10$ km, depth $z = -7$ km) contains 18 events with centroids that extend 500 m along strike but less than 20 m vertically; other examples are shown in Fig. 3. While some lineations appear to be seismogenic along their entire length (see, for example, Fig. 3c and d), others do not—at least on the basis of the 1984–97

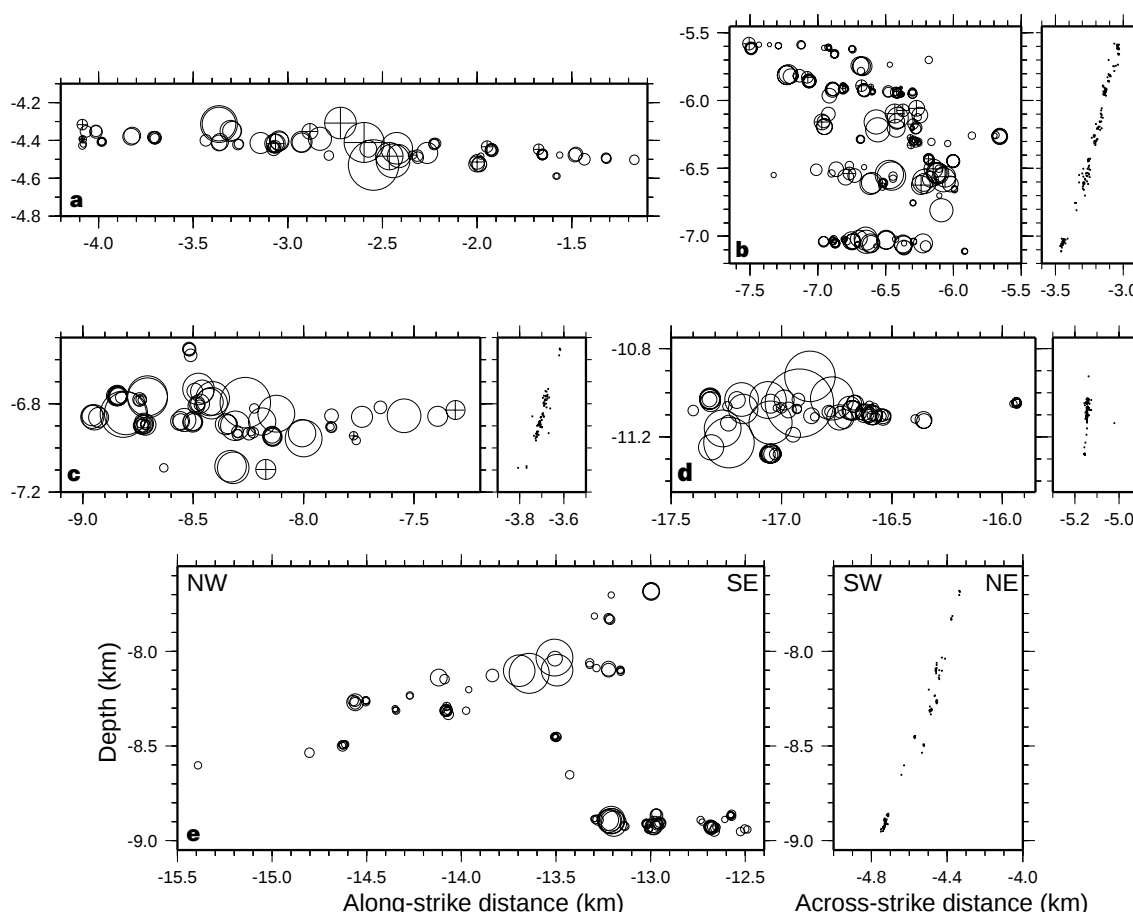


Figure 3 Vertical cross-sections of five multiplets, containing 75 to 85 events (**a**, **c** and **d**), 122 events (**e**), and 201 events (**b**). Circles with crosses indicate events that pre-date the Loma Prieta earthquake. Many repeating earthquakes cannot be distinguished because they plot atop one another. With the exception of **d**, these examples contain nearly all of the local NCSN catalogue seismicity. Several clusters of repeating earthquakes are located ~50 m deeper and to the southeast of pre-Loma Prieta earthquakes of the same size; this apparent offset is probably

an artefact of unmodelled velocity changes outside the source region following the Loma Prieta earthquake²³. The right-hand panels in **b–e** show across-strike sections. Most of the fault zone in the study area dips about 70° to the southwest, as seen in multiplets **b**, **c** and **e**; the fault zone steepens at shallower and greater depth, as in **d**. In **e**, the dip is well defined even in the 100-m-high lineation at 9 km depth; southeast of -13.5 km along-strike, the events over the full height of this multiplet are coplanar to within 20 m.

catalogue (for example, Fig. 3e). The multiplets contain numerous repeating earthquakes: that is, two or more events of essentially the same magnitude that occur in essentially the same spot. Repeating earthquakes in central California are well known^{19–22}, but previously the relative locations of neighbouring clusters of such events have not been determined on such a large scale. Other recent studies have imaged a few isolated horizontal lineations of microseismicity on strike-slip faults in California^{9,23} and Hawaii¹⁵.

Multiplets more than 1 km in length vary in thickness from less than 20 m to 100 m when viewed in a cross-section perpendicular to strike (Fig. 3b–e). Given the relocation errors, this does not distinguish them from planar. Many lineations have a resolvable dip despite their limited vertical extent; this is generally indistinguishable from that of the surrounding fault zone (Fig. 3c–e). Nearly all the multiplets consist of right-lateral strike-slip earthquakes; one exception (at $x = 9\text{--}10.5\text{ km}$, $z = -9\text{ km}$ in Fig. 2b) contains 33 thrust earthquakes that dip 45° to the northeast, indicative of compression perpendicular to the San Andreas fault. Visual inspection of all the waveforms for about half the multiplets indicates that, with very few exceptions, focal mechanisms within each multiplet are essentially identical. This is despite the fact that the multiplets contain numerous events classified in the NCSN catalogue as thrust or normal faulting earthquakes (about 15% of the total in a recent compilation of the best-constrained mechanisms¹⁷). Evidently, the published mechanisms are contaminated by errors in polarity determination.

Several criteria indicate that the relocations are robust. Because of excellent station coverage, many earthquake pairs have delay measurements at more than 60 stations, several of which have epicentral distances of less than one source depth, and over 100 combined P- and S-wave readings. Relocations of some linear multiplets using P-wave or S-wave measurements alone, North American or Pacific stations alone, or only stations for which the seismic rays leave the source region within 20° of horizontal or within 70° of vertical, yield similar images. We are confident that the linear fabric is not due to a spurious vertical ‘collapsing’ of the events for the following reasons. (1) Some multiplets consist of lineations at different depths (for example, Fig. 3b); (2) similar results are obtained when these lineations are relocated together and when they are relocated as separate multiplets placed at the average of their catalogue locations; (3) Dodge and Beroza²³ obtained a similar image of the multiplet in Fig. 3d using different multiplet selection, cross-correlation, and relocation techniques; and (4) the method places events in synthetic tests in their correct relative position. A useful method for assessing the solution visually is to plot all the delay measurements as a function of the angle between the vector connecting the events and the vector followed by the ray when leaving the source region (Fig. 4a and b). After accounting for the difference in origin times, this should yield a cosine curve with amplitude d/V , where d is the event separation and V is the seismic velocity in the source region²⁴. Figures 4a and b also clearly illustrate the velocity contrast across the fault in this region. Median residuals (the differences between the measured delays and those computed from the relocation) are 0.1 samples (1 ms) for event pairs separated by less than 25 m, consistent with a relative location accuracy of several metres. This is also consistent with the result that the earthquake centroids within several clusters of repeating events span less than 3 m. Even event pairs separated by 500 m have an average of 20 delay measurements and median residuals of less than 0.5 samples (Fig. 4c), consistent with the relative location accuracies (tens of metres) suggested by the thicknesses of the multiplets when viewed edge-on (Fig. 3b–e).

An important question concerns the nature of the 25% of the earthquakes that were not relocated. These tend to be very small or very large; the inset in Fig. 2a shows that their distribution is bimodal, with peaks at magnitudes 0.9 and 3.1. Waveforms of events near magnitude $M = 0.9$ have low signal-to-noise ratios

at most stations, whereas waveforms of events above $M = 3$ are severely clipped and cannot be cross-correlated except with nearby events of comparable size. Thus most of the unrelocated events may have been excluded by the multiplet selection process simply because of poor data quality.

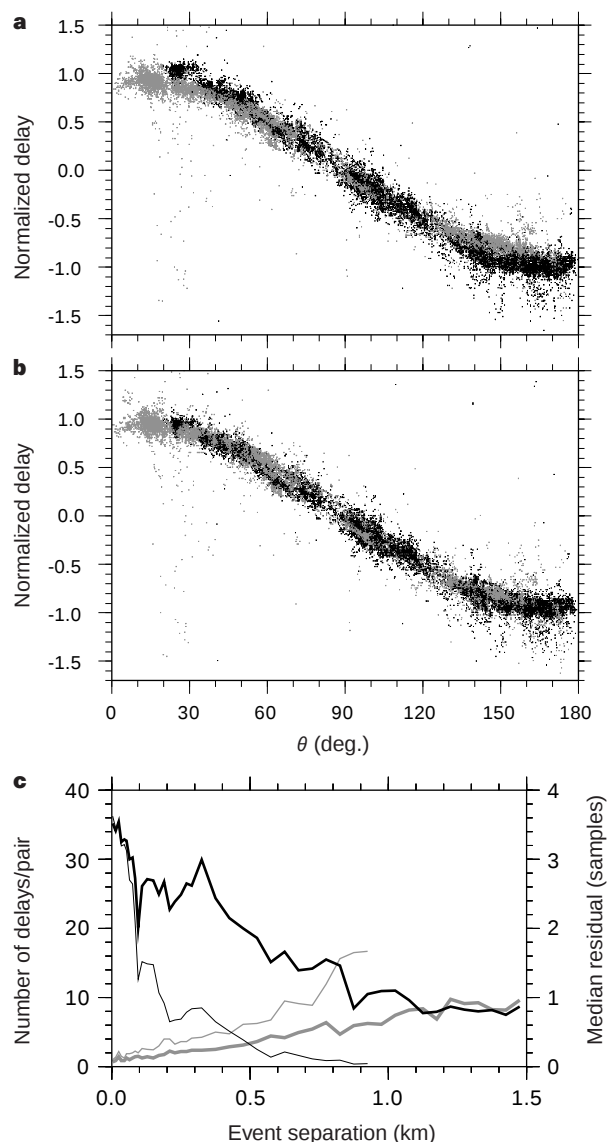


Figure 4 Solution statistics for the 122-event multiplet of Fig. 3e. **a, b**, Normalized delay measurements ($>24,000$) for all event pairs separated by 200–400 m. Delays at stations southwest of the San Andreas fault are shown in grey; those at stations northeast of the fault are shown in black. The travel-time delay between two events at station A is $(d/V)\cos\theta$, where d is the distance between the events, V is the seismic velocity in the source region, and θ is the angle between the vector connecting the events and the ray to station A. Here we plot as a function of θ the delay measurements normalized by d/V , where d and θ are taken from the relocation; for zero error in the model assumptions and delay measurements, this should yield a cosine of unit amplitude. In **a**, uniform velocities and a P/S velocity ratio of 1.75 are assumed. The curve defined by the Pacific-plate stations has a smaller amplitude, indicating a higher velocity. In **b**, events have been relocated after correcting for a 12% velocity contrast across the fault. This changes the locations only slightly, but lowers the median residual by 20%. **c**, Plot showing the decrease in the average number of delay measurements per earthquake pair (black curves) and the increase in median residual (grey curves) with increasing distance between the events (122 events implying 7,381 event pairs). P waves (thick lines) maintain their coherence for greater event separations than S waves (thin lines), perhaps because the S waves arrive within the P wavetrain.

A partial interpretation

Repeating earthquakes are generally interpreted as being produced by patches on a fault that undergo stick-slip failure as they are continually loaded by creep of the surroundings^{20,21,25}. This suggests that the fault is creeping between the lineations, and perhaps over a substantial portion of many lineations. One might still wish to consider the possibility that the apparently aseismic regions in Fig. 2b are locked, and that as the regional shear stress gradually increases, the lineations represent regions of low friction or low normal stress that fail repeatedly before the intervening regions fail once. (Regions of low normal stress could arise, for example, as a result of geometric mismatches across the fault surface.) Eventually, however, the intervening regions must fail. In some cases this failure could be signalled by the larger, unrelocated earthquakes shown in Fig. 2a. We do not think that this is a general explanation for the apparently aseismic regions, however, because in some areas there appear to be no candidate events of significant magnitude. Furthermore, because the repeating earthquake sequences in this data set began (or became much more active) following the Loma Prieta earthquake, if the apparently aseismic regions were locked one would have to envision a mechanism whereby the Loma Prieta earthquake increased the regional stressing rate. Schaff *et al.*²⁵, on the other hand, found that the recurrence intervals in several clusters of repeating earthquakes between $x = -17$ km and $x = -19$ km in Fig. 2b increased logarithmically with time following the Loma Prieta earthquake, consistent with creep on the surrounding fault being induced by that earthquake. Thus we infer that the aseismic regions are undergoing creep.

An important question is whether the microseismic lineations are structural or compositional in origin. An example of the former would be some geometric feature that tends to be aligned in the direction of slip; an example of the latter would be strata of unique mineralogy, fluid content or metamorphic reactions that favour stick-slip as opposed to stable sliding. Based on the California earthquakes alone the answer might be ambiguous, but in conjunction with data from faults of different orientations in Hawaii (Fig. 5a) and Alaska it appears that the fabric is structural. A

multiplet 5 km west of Kilauea caldera, Hawaii, defines a fault plane that strikes north-south and dips 60° to the west; when viewed face-on, the microearthquakes are concentrated in bands aligned in the dip direction (Fig. 5b). On the basis of the dip, it seems most likely that this is a normal fault (the focal mechanisms are poorly constrained but would allow this); if this is the case, then the lineations are parallel to the slip direction. Two multiplets southeast of the caldera define roughly horizontal planes at about 8 km depth; these are inferred to lie along the basal decollement beneath Kilauea's south flank⁵. In this case, the direction of slip is well constrained by geodetic data, which show the upper fault block to be creeping at rates of 10–20 cm yr⁻¹ to the southeast^{26,27}. As with the multiplet west of the caldera, the station coverage here is much less dense than in central California and the earthquake locations are not as well constrained; nonetheless, the evidence seems quite strong that these faults exhibit streaks of microseismicity aligned nearly in the slip direction (Fig. 5c, d). On a rather larger scale, Abers *et al.*²⁸ also identified a few bands of seismicity parallel to slip on the subduction interface beneath the Shumagin islands, south-east Alaska.

There are numerous candidate mechanisms for generating linear fabrics within fault zones. Field observations indicate that fault surfaces are corrugated on a variety of scales, with the lineations running parallel to the slip direction²⁹. Faults are also segmented on a variety of scales¹, and with large-scale slip those segment boundaries that parallel the slip direction might be better preserved than those that are oblique. Other possibilities are that blocks of resistant host rock are smeared out along the fault zone as slip proceeds, or that resistant protrusions on one side of the fault plough through a gouge zone, leaving a linear fabric in their wake²⁸.

The mechanisms by which any of these structures might give rise to a linear trend of earthquakes are potentially varied and complex. Earthquake generation is an instability requiring that as slip accelerates, the resistance to sliding decreases faster than the stress driving that slip³⁰. This balance is affected by both the intrinsic frictional properties of the fault-zone materials and the geometry of the fault surface. At high temperature, friction increases with sliding

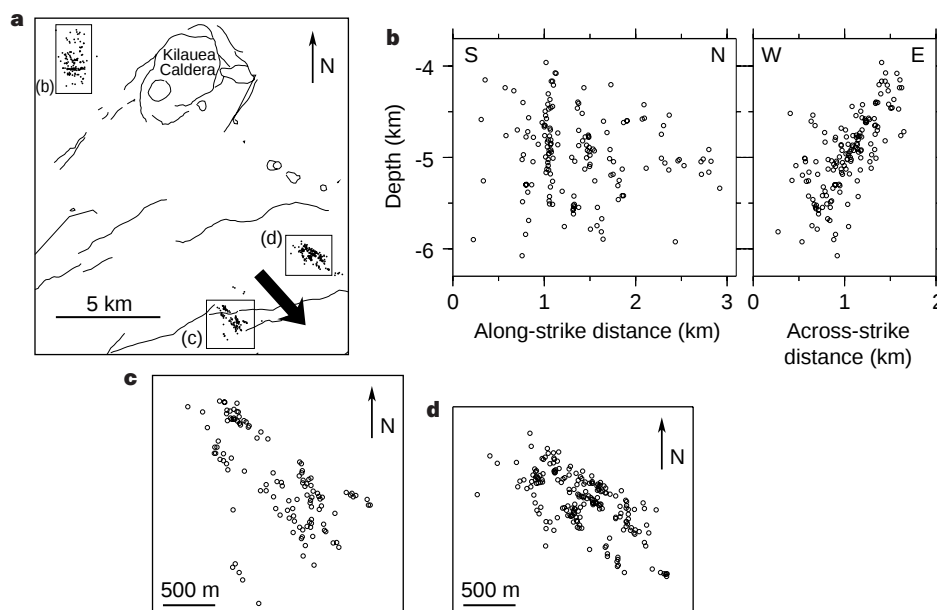


Figure 5 Relocated multiplets from Kilauea volcano, showing bands of microearthquakes nearly parallel to the inferred slip direction. **a**, Map of Kilauea volcano showing three relocated multiplets (labelled b, c and d) extracted from the 1987–88 earthquake catalogue. Arrow shows the approximate direction of south flank motion determined from geodetic data. **b**, Vertical cross-sections of multiplet b, located 5 km west of Kilauea caldera, which strikes north–south and

dips ~60° to the west. The station coverage is less dense than in central California, and thus the vertical errors are considerably larger. **c**, **d**, Map views of multiplets c and d, located respectively 10 km SSE and SE of Kilauea caldera, inferred to lie on the near-horizontal decollement beneath the south flank at a depth of ~8 km.

velocity; this is the standard explanation for the lack of seismicity below 10–15 km depth along most of the San Andreas fault (see, for example, Fig. 2). Low normal stress and the presence of unconsolidated gouge is the standard explanation for the lack of seismicity at depths shallower than 3–4 km (ref. 1). Factors that could promote stick–slip behaviour at intermediate depths on an otherwise creeping fault include (1) high normal stress³⁰, (2) a normal stress that decreases with slip, as a result of a non-planar fault surface³¹, and (3) a reduced gouge thickness³².

Given the number of possible mechanisms for generating linear fabrics within fault zones, and the number of ways in which this fabric might interact with the frictional constitutive law to produce earthquakes, it is difficult to pinpoint a single cause for the observed lineations in microseismicity. Corrugations of the fault surface or fault offsets could generate linear regions of high normal stress, depending on the stress tensor, and if they are not perfectly aligned with the current slip vector they could generate linear regions where the normal stress decreased with slip. Blocks of rock within the gouge that resist comminution could promote earthquake generation by providing rock-on-rock slip surfaces or reducing the gouge thickness. At present, all that can be said of the underlying structural heterogeneity is that it cannot be resolved by the current relocation technique. That is, many of the lineations have a resolvable dip that is indistinguishable from that of the surrounding fault zone (for example, Fig. 3c–e). In addition, from first motion polarities the focal mechanisms of vertically adjacent lineations (such as those in Fig. 3b) appear to be identical. Thus corrugations of the fault surface, if they exist, must produce only very subtle changes in dip; similarly, resistant blocks within the fault zone, if they are the cause of the earthquakes, must produce slip surfaces with little variation in orientation.

Patterns in time and space

With a few exceptions, there is no obvious temporal migration of seismicity within most multiplets. Nonetheless, large data sets

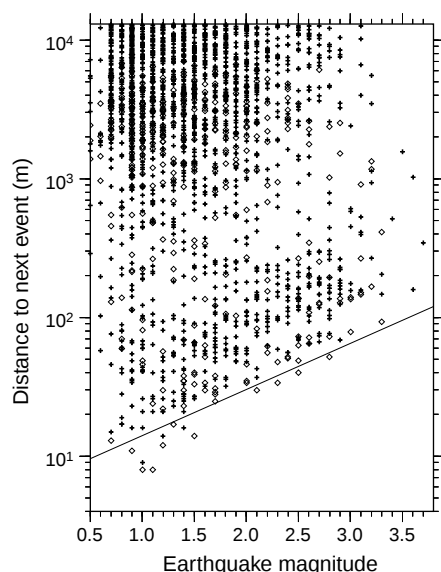


Figure 6 Plot of the distance to the second earthquake as a function of the magnitude of the first, for each consecutive pair of relocated earthquakes. Diamonds indicate that the vertical separation between the events exceeds the horizontal separation; crosses indicate the reverse. The diagonal line is the estimate of earthquake radius derived from the moment–magnitude relation of Abercrombie¹⁸, assuming a stress drop of 10 MPa. Abercrombie³⁵ found that most events in this magnitude range had stress drops of 1–10 MPa; hence our choice of 3 MPa for determining the circle sizes in Figs 2 and 3.

such as this are likely to contain significant new information about rupture interaction, primarily because the errors in relative location are smaller than the rupture dimensions. Figure 6 shows, for each consecutive pair of relocated earthquakes, the distance to the second event as a function of the magnitude of the first. Because earthquakes reduce the shear stress on the ruptured area but increase the shear stress beyond the rupture perimeter, aftershocks are most likely to occur some short distance beyond the perimeter of the previous event. If earthquake occurrence were random, the density of points in Fig. 6 would increase with distance because of the logarithmic distance scale, a trend observed in the upper half of the figure. The increase in density below roughly 100 m for $M = 1$ events and 1 km for $M = 3.5$ events can thus be attributed to (smaller or larger) aftershocks. A striking feature of this figure is the absence of aftershocks closer than about 10 m to an $M = 1$ event and 100 m to an $M = 3.5$ event; we interpret this cut-off to be an indication of earthquake size. In support of this, the diagonal line in Fig. 6 indicates the earthquake radius derived from the moment–magnitude relation of Abercrombie¹⁸, assuming a stress drop of 10 MPa (reducing the stress drop to 1 MPa shifts the line upward by a factor of $10^{1/3} \approx 2$ but leaves its slope unchanged). While this behaviour is not unexpected³³, it has been difficult to observe for small earthquakes because of inadequate catalogue locations (event pairs separated by less than 20 m in Fig. 6 have a median separation of 700 m in the NCSN catalogue). In addition, such a simple mainshock/aftershock relationship is often not seen following large earthquakes, where many aftershocks seem to occur on the portion of the fault that underwent slip during the mainshock (presumably because aftershocks may lie on secondary faults adjacent to the mainshock, or in regions of large gradients in mainshock slip; see, for example, ref. 34).

How common a feature?

Our observation of similar lineations on a strike–slip fault, a (presumed) steep normal fault, and a nearly horizontal thrust fault suggests that streaks of microearthquakes aligned in the slip direction are a common characteristic of creeping faults. Because faults that undergo most of their slip aseismically appear to be a minority¹, this need not imply that such lineations are a common characteristic of faults generally. However, it is worth noting that until recently the San Andreas fault northwest of San Juan Bautista had generally been referred to as locked. So how common are microseismic lineations? Shearer⁷ did not observe them when using a related technique to relocate nearly 600 aftershocks of the 1986 Whittier Narrows earthquake in California, and we could identify only a few small multiplets in a localized subset of 1,000 aftershocks of the 1994 Northridge earthquake (also in California). One possibility is simply that lineations are not present on faults that undergo most of their slip seismically. An alternative interpretation, however, is that these data sets cover too short a time period for any spatial structure in the microseismicity to become apparent. Focal mechanisms of significant aftershock sequences in particular are commonly quite diverse, so an apparently high earthquake density might indicate that a large number of secondary faults have been activated, and need not imply a high density of earthquakes on a single fault. If this is the case, then lineations in these regions might only become apparent given enough time and earthquake activity. Put another way, one can ask whether the statement “microseismic lineations are a common feature of creeping faults” can be expanded to “microseismic lineations are a common feature of faults when they creep”, even if the fault is dominantly seismogenic. The latter interpretation is consistent with the pronounced fabric on the San Andreas fault northwest of San Juan Bautista, which slipped by 2 m during the 1906 earthquake and which appears to have undergone accelerated creep (including a 30-fold increase in the seismic moment and earthquake production rates) following the Loma Prieta earthquake. □

Methods

The measure of similarity of two traces is the average coherency, in the frequency range 4–15 Hz, of a 1.28-s (128 time samples) window containing the P wavetrain. Defining a multiplet requires choosing a minimum average coherency for two seismograms to be considered similar (typically 95–85%), a minimum number of stations at which this coherency is exceeded for two events to be considered similar (typically 5–10), and the minimum number of other events in the multiplet to which each member must be similar (typically 1–5). The precise criteria vary depending on the local earthquake density and station coverage. Delays within each multiplet are determined by cross-spectral analysis¹⁹, using 128-sample windows and a frequency range of 4–26 Hz. We use vertical-component stations for P-wave measurements and vertical and horizontal components for S-wave measurements. We begin with rather strict multiplet selection criteria and progressively relax these until the multiplets become too large (more than ~3 km across) or until events are included that have median residuals significantly larger than the median residuals obtained using the previous selection criteria.

Received 10 March; accepted 1 July 1999.

1. Scholz, C. H. *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, 1990).
2. Allen, C. R., St Amand, P., Richter, C. F. & Nordquist, J. M. Relationship between seismicity and geologic structure in the southern California region. *Bull. Seismol. Soc. Am.* **55**, 753–797 (1965).
3. Wallace, R. E. & Morris, H. T. Characteristics of faults and shear zones in deep mines. *Pure Appl. Geophys.* **124**, 107–125 (1986).
4. Chester, F. M., Evans, J. P. & Biegel, R. L. Internal structure and weakening mechanisms of the San Andreas fault. *J. Geophys. Res.* **98**, 771–786 (1993).
5. Got, J.-L., Fréchet, J. & Klein, F. Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea. *J. Geophys. Res.* **99**, 15375–15386 (1994).
6. Dodge, D. A., Beroza, G. C. & Ellsworth, W. L. Foreshock sequence of the 1992 Landers, California earthquake and its implications for earthquake nucleation. *J. Geophys. Res.* **100**, 9865–9880 (1995).
7. Shearer, P. M. Improving local earthquake locations using the L1 norm and waveform cross-correlation: Application to the Whittier Narrows, California, aftershock sequence. *J. Geophys. Res.* **102**, 8269–8283 (1997).
8. Lees, J. M. Multiplet analysis at Coso Geothermal. *Bull. Seismol. Soc. Am.* **88**, 1127–1143 (1998).
9. Waldhauser, F., Ellsworth, W. L. & Cole, A. Slip-parallel seismic lineations on the northern Hayward fault, California. *Geophys. Res. Lett.* (submitted).
10. Gladwin, M. T., Gwyther, R. L., Hart, R. H. G. & Breckenridge, K. S. Measurements of the strain field associated with episodic creep events on the San Andreas fault at San Juan Bautista, California. *J. Geophys. Res.* **99**, 4559–4565 (1994).
11. Linde, A. T., Gladwin, M. T., Johnson, M. J. S., Gwyther, R. L. & Bilham, R. G. A slow earthquake sequence on the San Andreas Fault. *Nature* **383**, 65–68 (1996).
12. Johnston, M. J. S. Near-silent earthquakes on the San Andreas fault system. *Eos* **78**, F156 (1997).
13. Thatcher, W., Marshall, G. & Lisowski, M. Resolution of fault slip along the 470-km-long rupture of the great 1906 San Francisco earthquake and its implications. *J. Geophys. Res.* **102**, 5353–5367 (1997).
14. VanDecar, J. C. & Crosson, R. S. Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares. *Bull. Seismol. Soc. Am.* **80**, 159–169 (1990).

15. Gillard, D., Rubin, A. M. & Okubo, P. Highly concentrated seismicity caused by deformation of Kilauea's deep magma system. *Nature* **384**, 343–346 (1996).
16. Thurber, C. H. *et al.* Two-dimensional seismic images of the San Andreas Fault in the Northern Gabilan Range, central California: Evidence for fluids in the fault zone. *Geophys. Res. Lett.* **24**, 1591–1594 (1997).
17. Eberhart-Phillips, D. & Michael, A. J. Seismotectonics of the Loma Prieta, California region determined from three-dimensional V_p , V_p/V_s , and seismicity. *J. Geophys. Res.* **103**, 21099–21120 (1998).
18. Abercrombie, R. E. The magnitude-frequency distribution of earthquakes recorded with deep seismometers at Cajon Pass, southern California. *Tectonophysics* **261**, 1–7 (1996).
19. Poupinet, G., Ellsworth, W. L. & Frechet, J. Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. *J. Geophys. Res.* **102**, 5719–5731 (1984).
20. Vidale, J. E., Ellsworth, W. L., Cole, A. & Marone, C. Variations in rupture process with recurrence interval in a repeated small earthquake. *Nature* **368**, 624–626 (1994).
21. Nadeau, R. M., Foxall, W. & McEvilly, T. V. Clustering and periodic recurrence of microearthquakes on the San Andreas Fault at Parkfield, California. *Science* **267**, 503–507 (1995).
22. Beroza, G. C., Cole, A. T. & Ellsworth, W. L. Stability of coda wave attenuation during the Loma Prieta, California, earthquake sequence. *J. Geophys. Res.* **100**, 3977–3987 (1995).
23. Dodge, D. A. & Beroza, G. C. Source array analysis of coda waves near the 1989 Loma Prieta, California, mainshock: Implications for the mechanism of coseismic velocity changes. *J. Geophys. Res.* **102**, 24437–24458 (1997).
24. Got, J.-L. & Coutant, F. Anisotropic scattering and travel time delay analysis in Kilauea volcano, Hawaii, earthquake coda waves. *J. Geophys. Res.* **102**, 8397–8410 (1997).
25. Schaff, D. P., Beroza, G. C. & Shaw, B. E. Postseismic response of repeating aftershocks. *Geophys. Res. Lett.* **25**, 4549–4552 (1998).
26. Delaney, P. T., Miklius, A., Arnadottir, T., Okamura, A. T. & Sato, M. Motion of Kilauea Volcano during sustained eruption from the Puu'Oo and Kupaianaha vents, 1983–1991. *J. Geophys. Res.* **98**, 17801–17820 (1993).
27. Owen, S. *et al.* Rapid deformation of the south flank of Kilauea Volcano, Hawaii. *Science* **267**, 1328–1332 (1995).
28. Abers, G. A., Hu, X. & Sykes, L. R. Source scaling of earthquakes in the Shumagin region, Alaska: Time-domain inversions of regional waveforms. *Geophys. J. Int.* **123**, 41–58 (1995).
29. Power, W. L. & Tullis, T. E. The contact between opposing fault surfaces at Dixie Valley, Nevada, and implications for fault mechanics. *J. Geophys. Res.* **97**, 15425–15436 (1992).
30. Dieterich, J. H. Earthquake nucleation on faults with rate and state-dependent friction. *Tectonophysics* **211**, 149–178 (1992).
31. Dieterich, J. H. & Linker, M. F. Fault stability under conditions of variable normal stress. *Geophys. Res. Lett.* **19**, 1691–1694 (1992).
32. Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* **26**, 643–696 (1998).
33. Ellsworth, W. L. in *Urban Disaster Mitigation* (eds Cheng, F. Y. & Sheu, M.-S.) 1–14 (Elsevier, Tarrytown, New York, 1995).
34. Beroza, G. & Zoback, M. L. Mechanism diversity of the Loma Prieta aftershocks and the mechanics of mainshock-aftershock interaction. *Science* **259**, 210–213 (1993).
35. Abercrombie, R. E. Earthquake source scaling relationships from -1 to $5 M_L$ using seismograms recorded at 2.5-km depth. *J. Geophys. Res.* **100**, 24015–24036 (1995).

Acknowledgements. We thank R. Abercrombie for discussions, F. Waldhauser, W. L. Ellsworth and A. Cole for a preprint of their Hayward fault relocations, and F. A. Dahlen, J. Lees and P. Shearer for comments. The figures were produced using the GMT software provided by P. Wessel and W. Smith. This work was supported by the US NSF and the US Geological Survey.

Correspondence and requests for materials should be addressed to A.M.R. (e-mail: arubin@princeton.edu).

Photon-stimulated desorption as a substantial source of sodium in the lunar atmosphere

B. V. Yakshinskiy & T. E. Madey

Department of Physics and Astronomy, and Laboratory for Surface Modification, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854-8019, USA

Mercury and the Moon both have tenuous atmospheres that contain atomic sodium and potassium. These chemicals must be continuously resupplied, as neither body can retain the atoms for more than a few hours (refs 1–6). The mechanisms proposed to explain the resupply include sputtering of the surface by the solar wind^{3,4}, micrometeorite impacts⁵, thermal desorption^{6–8} and photon-stimulated desorption^{6–10}. But there are few data and no general agreement about which processes dominate^{5,10–14}. Here we report laboratory studies of photon-stimulated desorption of sodium from surfaces that simulate lunar silicates. We find that bombardment of such surfaces at temperatures of ~ 250 K by ultraviolet photons (wavelength $\lambda < 300$ nm) causes very efficient desorption of sodium atoms, induced by electronic excitations rather than by thermal processes or momentum transfer. The flux at the lunar surface of ultraviolet photons from the Sun is sufficient to ensure that photon-stimulated desorption of sodium contributes substantially to the Moon's atmosphere. On Mercury, solar heating of the surface implies that thermal desorption⁷ will also be an important source of atmospheric sodium.

We made measurements in an ultrahigh-vacuum apparatus, with base pressures $< 1 \times 10^{-10}$ torr (ref. 15), and used X-ray photoelectron spectroscopy, temperature-programmed desorption and low-energy ion scattering for surface characterization. Amorphous, stoichiometric SiO_2 films ~ 10 nm thick were prepared by thermal evaporation of silicon in an oxygen atmosphere (1×10^{-5} torr), onto the surface of a Re(0001) single crystal¹⁶. These films simulate lunar silicates (SiO_2 is the dominant component of the lunar surface). Sodium was deposited from an evaporation source, and coverages were measured by X-ray photoelectron spectroscopy and temperature-programmed desorption (ref. 7, and B.V.Y., V. N. Ageev and T.E.M., manuscript in preparation). The measurement scheme for photon-stimulated desorption (PSD) of Na atoms includes a mechanically chopped photon source (500-W Hg arc lamp with filters), a highly sensitive detector of Na based on surface ionization, and the use of a time-of-flight technique. For companion studies of electron-stimulated desorption (ESD), an efficient pulsed low-energy electron source was used together with the Na detector.

There are intimate connections between ESD and PSD¹⁷; in general, similar electronic processes cause desorption of atoms via electron or photon excitation. The electronic excitations accessible using photons can also be excited by electrons of comparable energy, and a tunable electron gun (0–200 eV) has a much wider energy range than a laboratory photon source (Hg arc lamp, $E_{\text{max}} \approx 5$ eV). Thus, we measure low-energy desorption thresholds and cross-sections using PSD, and probe cross-sections and mechanisms over a wider energy range using ESD.

Figure 1a shows the cross-section (cm^2) for Na atom desorption via PSD as a function of photon energy $h\nu$. Visible and near-ultraviolet photons ($\lambda > 300$ nm, $h\nu < 4$ eV) cause little or no detectable desorption of Na. For ultraviolet photons with $h\nu > 4$ eV, the cross-section rises; at $h\nu \approx 5$ eV, the PSD cross-section is $\sim (3 \pm 1) \times 10^{-20} \text{ cm}^2$. The ESD cross-section for atomic Na as a function of electron energy E_e is shown in Fig. 1b. We find that (1) the initial threshold is at ~ 4 eV, the same as the PSD threshold,

(2) the desorption cross-sections at ~ 5 eV have similar magnitudes for ESD and PSD and (3) there is a resonance-like feature at ~ 11 eV in ESD of Na. Similar features are seen for all Na coverages less than 1 monolayer (ML) at a substrate temperature of ~ 250 K for both PSD and ESD. The yields and thresholds are also substantially the same for a Na/ SiO_2 surface that has been heavily bombarded by ions (1-keV He^+), which simulates damage by the solar wind. The threshold for desorption of ionic Na^+ via ESD is ~ 25 eV, significantly higher than for neutral Na.

The velocity distribution for ESD of neutral Na is shown in Figure 2. The peak in the distributions is at $1,000 \text{ m s}^{-1}$; the peak in the corresponding kinetic-energy distribution is at ~ 0.1 eV. The velocities (energies) are clearly non-thermal (substrate at 250 K), and have values consistent with the 'hot' components of the lunar and mercurian atmospheres^{11,18–20}.

We interpret the PSD and ESD data of Fig. 1a, b on the basis of charge-transfer excitations^{21,22} from SiO_2 to the unoccupied Na 3s orbital (Fig. 3), so that adsorbed Na^+ is neutralized and desorbs. Na is adsorbed in ionic form on SiO_2 for coverages < 1 ML (B.V.Y., V. N. Ageev and T.E.M., manuscript in preparation). The radius of ionic Na^+ (~ 1.0 Å) is smaller than that of neutral Na^0 (~ 1.9 Å). When an electron is attached to Na^+ to make Na^0 , the neutral adsorbed atom is at a smaller atom-surface separation than the equilibrium value; this is a repulsive configuration which can lead to desorption of a non-thermal Na atom. Desorption of Na involves a competition

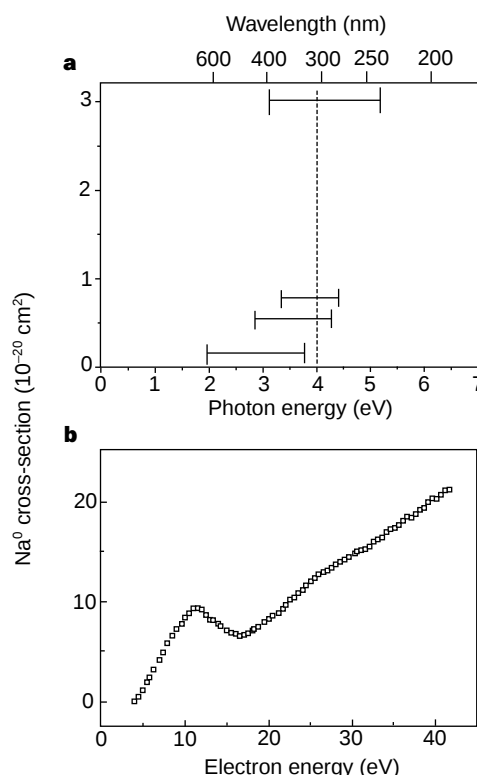


Figure 1 Photon-stimulated desorption (PSD) and electron-stimulated desorption (ESD) of Na from SiO_2 at 250 K. The cross-section for Na atom desorption is plotted as a function of primary excitation energy. **a**, PSD cross-section versus photon energy $h\nu$. The horizontal bars represent the effective bandwidth of the filters; short-wavelength photons (that is, high-energy photons > 4 eV), as indicated by the dashed line) are most effective in causing desorption. **b**, ESD cross-section versus electron energy E_e . The average incident electron current density is $\sim 1 \times 10^{-6} \text{ A cm}^{-2}$; E_e is corrected for work functions. For both **a** and **b**, the Na coverage is 0.22 ML. The detector for neutral Na is based on surface ionization of Na on hot ($\sim 1,600$ K) iridium ribbons. The Na^+ ions that desorb from the Ir ribbons are detected using a channeltron multiplier. In each case, the cross-section is determined at a fixed energy from the decay of Na desorption rate with time during extended irradiation of the sample (~ 2 h).

between electron-excited or photon-excited electron transfer to form Na^0 , and the lifetime of Na^0 in a repulsive configuration. On SiO_2 , because the Na 3s level is in the bandgap region between the Fermi level and the conduction-band minimum, the lifetime of the neutral atom is long enough for efficient desorption. In contrast, the ESD Na yields from oxidized surfaces of tungsten and molybdenum are negligible for $E_e < 25$ eV (refs 23, 24): the Fermi level is near the conduction-band minimum, and the neutral lifetime is too short for efficient desorption.

The sodium source rate needed to maintain the observed column density of $\sim 10^9$ atoms cm^{-2} above the lunar surface (that is, the desorption rate of Na from the lunar surface) is variously

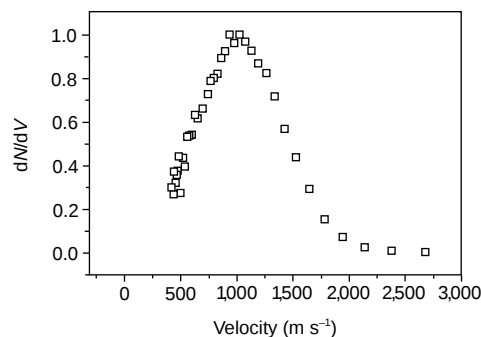


Figure 2 Velocity distribution for ESD of neutral Na from SiO_2 (0.22 ML); electron energy $E_e = 200$ eV. The electron source is pulsed, and Na atoms that desorb from SiO_2 have flight times of ~ 50 μs to the detector Ir ribbons (see Fig. 1 legend). The time-of-flight technique is highly sensitive for detection of Na atoms, with low background signal. dN is the number of desorbing atoms having velocities within the range $V + dV$.

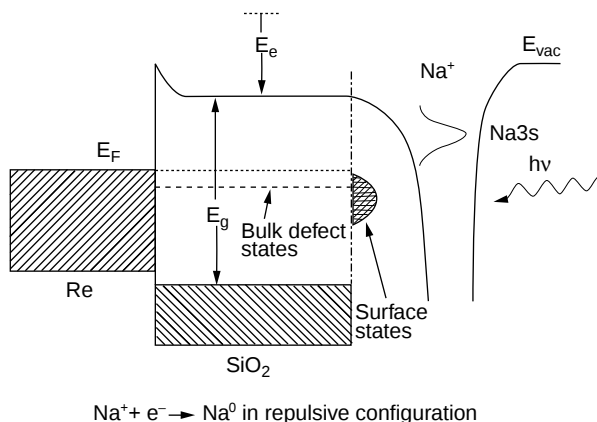


Figure 3 Schematic model for photon- or electron-stimulated desorption of neutral Na from SiO_2 . PSD or ESD occurs via photon-excited or electron-excited charge transfer to Na^+ . Here, $h\nu$ is the incident photon energy, E_e is the incident electron energy, E_{vac} is the vacuum energy level, E_F is the Fermi energy, and E_g is the bandgap energy of amorphous SiO_2 (~ 8 – 9 eV). The unfilled Na 3s level is also indicated. The charge-transfer electrons may be produced by electron bombardment (ESD) or by photo-excitation of substrate electrons (PSD). The threshold at ~ 4 eV is associated with electronic excitations from SiO_2 surface states, or from SiO_2 defect states¹⁶, to the unfilled Na 3s level. (The surface states are presumed to be localized states formed by the trapping of the Na valence electron on adsorption of Na^+). In Fig. 1b, the resonance-like peak at ~ 11 eV may be associated with bandgap excitations, broadened by band-bending effects. Higher-energy excitations (including the shoulder at ~ 25 eV) may be due to multielectronic excitations as well as excitations initiated by shallow core hole ionization (for example, O 2s). The high Na^0 yields at low electron or photon excitation energies appear to be related to a very small barrier between the conduction-band minimum of SiO_2 and Na^+ : this leads to a high probability for electron capture by Na^+ of a transient electron propagating in the conduction band.

estimated^{11,25} to be in the range $\sim 5 \times 10^3$ to $\sim 2 \times 10^6$ atoms $\text{cm}^{-2} \text{s}^{-1}$. The different values are mainly based on different assumptions about the escape dynamics of Na. We can estimate the PSD flux $\Phi_{\text{Na}}^{\text{PSD}}$ of Na from the lunar surface induced by ultraviolet photons, based on our measurements and some assumptions. The PSD cross-section Q is $\sim (3 \pm 1) \times 10^{-20} \text{cm}^2$ at $h\nu \approx 5$ eV. The solar photon flux F_{ph} (ref. 26) at the lunar surface is $\sim 2 \times 10^{14}$ photons $\text{cm}^{-2} \text{s}^{-1}$ at $h\nu \geq 5$ eV, and decreases very quickly with increasing energy; the flux is $\sim 10^{12}$ photons $\text{cm}^{-2} \text{s}^{-1}$ at $h\nu > 8$ eV, above the bandgap energy. If we assume the Na surface coverage σ to be $\sim 3 \times 10^{12} \text{cm}^{-2}$, consistent with the lunar average bulk concentration of Na (ref. 5), we find

$$\Phi_{\text{Na}}^{\text{PSD}} \approx \frac{1}{4} F_{\text{ph}} Q \sigma \approx 4 \times 10^6 \text{ atoms cm}^{-2} \text{s}^{-1}$$

where the factor of 1/4 gives a surface-averaged value. Considering the uncertainties and variations in the lunar surface composition and the concentration of Na, the computed value is in good agreement with the upper limit of Na source rates indicated above.

For comparison, we estimate the possible role of ESD in the production of Na, due to bombardment of the lunar surface by electrons from the solar plasma. Based on the average electron flux of $\sim 4 \times 10^8 \text{cm}^{-2} \text{s}^{-1}$ with a mean temperature of $\sim 1.4 \times 10^5$ K (ref. 27), the ESD cross-section versus electron energy E_e in Fig. 1b, and the same Na surface coverage as above, the ESD flux $\Phi_{\text{Na}}^{\text{ESD}}$ is ~ 100 atoms $\text{cm}^{-2} \text{s}^{-1}$. Although there are considerable uncertainties in the parameters as applied to the lunar surface, $\Phi_{\text{Na}}^{\text{ESD}}$ is considerably less than both $\Phi_{\text{Na}}^{\text{PSD}}$ and the range of Na source rates from the lunar surface.

We now consider how the charge-transfer process occurs in PSD and ESD of Na. In PSD by ultraviolet photons, desorption can proceed via a direct excitation (that is, by photo-excitation of an electron from a surface state to neutralize surface Na^+), or via an indirect process²⁸. In the latter case, the photons excite 'hot' electrons to the conduction band from defect states in amorphous SiO_2 , or across the bandgap from the valence band; these electrons can propagate to the surface and neutralize surface Na^+ , which then desorbs with a non-thermal velocity distribution. Thus, the PSD/ESD electronic excitation processes are similar, and the cross-sections for Na desorption near threshold are of the same order of magnitude. In the present case, the direct electron flux from the solar wind is clearly too small to cause significant desorption of 'hot' Na, but the solar photon flux appears to be more than sufficient.

These measurements provide strong scientific support and a rationale for the arguments^{6–10}, discussed recently by Mendillo *et al.*²⁵, that a PSD process is important in production of the lunar Na atmosphere. Other processes^{3–5,25} also contribute; for example, a micrometeor source may be significant when associated with a meteor shower²⁹. As micrometeor impacts can change the local surface concentration of Na, some of the observed temporal and spatial variations in atmospheric Na concentrations may be partly due to surface-coverage-dependent PSD fluxes (as well as being affected by rates of impact vaporization of Na). In the case of Mercury, thermal desorption of Na is believed to contribute to the Na atmosphere because of the high day temperatures (~ 700 K at the sub-solar point¹¹), but PSD is also likely to be important, especially at other latitudes. □

Received 20 January; accepted 30 June 1999.

- Potter, A. E. Jr & Morgan, T. H. Discovery of sodium in the atmosphere of Mercury. *Science* **229**, 651–653 (1985).
- Potter, A. E. Jr & Morgan, T. H. Discovery of sodium and potassium vapor in the atmosphere of the Moon. *Science* **241**, 675–680 (1988).
- Potter, A. E. Jr & Morgan, T. H. Variation of the lunar sodium emission intensity with phase angle. *Geophys. Res. Lett.* **21**, 2263–2266 (1994).
- Potter, A. E. Jr & Morgan, T. H. Coronagraphic observation of the lunar sodium surface. *J. Geophys. Res.* **103**, 8581–8586 (1998).
- Morgan, T. H. & Shemansky, D. E. Limits to the lunar atmosphere. *J. Geophys. Res.* **96**, 1351–1367 (1991).
- McGrath, M. A., Johnson, R. E. & Lanzerotti, L. J. Sputtering of sodium on the planet Mercury. *Nature* **323**, 694–696 (1986).

7. Madey, T. E., Yakshinskiy, B. V., Ageev, V. N. & Johnson, R. E. Desorption of alkali atoms and ions from oxide surfaces: Relevance to origins of Na and K in atmospheres of Mercury and the Moon. *J. Geophys. Res.* **103**, 5873–5887 (1998).
8. Sprague, A. L. *et al.* Distribution and abundance of sodium in Mercury's atmosphere, 1985–1988. *Icarus* **129**, 506–527 (1997).
9. Ip, W. The sodium exosphere and magnetosphere of Mercury. *Geophys. Res. Lett.* **13**, 423–426 (1986).
10. Mendillo, M. & Baumgardner, J. Constraints on the origin of the Moon's atmosphere from observations during a lunar eclipse. *Nature* **377**, 404–406 (1995).
11. Hunten, D. M. & Sprague, A. L. Origin and character of the lunar and Mercurian atmospheres. *Adv. Space Res.* **19**, 1551–1560 (1997).
12. Sprague, A. L., Kozlowski, R. W. H. & Hunten, D. M. Caloris Basin: An enhanced source for potassium in Mercury's atmosphere. *Science* **249**, 1140–1143 (1990).
13. Sprague, A. L., Kozlowski, R. W. H., Hunten, D. M., Wells, W. K. & Grosse, F. A. The sodium and potassium atmosphere of the Moon and its interaction with the surface. *Icarus* **96**, 7–42 (1992).
14. Mendillo, M., Emery, J. & Flynn, B. Modeling the Moon's extended sodium cloud as a tool for investigating sources of transient atmospheres. *Adv. Space Res.* **19**, 1557–1560 (1997).
15. Maschhoff, B. L., Pan, J.-M. & Madey, T. E. Medium-energy backscattered electron diffraction studies of TiO₂(110): relation to surface structure. *Surf. Sci.* **259**, 190–206 (1991).
16. Xu, X. & Goodman, D. W. The preparation and characterization of ultra-thin silicon dioxide films on a Mo (110) surface. *Surf. Sci.* **282**, 323–332 (1993).
17. Knotek, M. L. in *Desorption Induced by Electronic Transitions* (eds. Talk, N. H., Traum, M. M., Tully, J. C. & Madey, T. E.) 139–155 (Springer Ser. in Chemical Physics, Vol. 24, Springer, Berlin, 1983).
18. Potter, A. E. Jr & Morgan, T. H. Variation of sodium on Mercury with solar radiation pressure. *Icarus* **71**, 472–477 (1987).
19. Kozlowski, R. W. H., Sprague, A. L. & Hunten, D. M. Observation of potassium in the tenuous lunar atmosphere. *Geophys. Res. Lett.* **17**, 2253–2256 (1990).
20. Flynn, B. & Mendillo, M. Simulations of the lunar sodium atmosphere. *J. Geophys. Res.* **100**, 23271–23278 (1995).
21. Hellsing, B., Chakarov, D. V., Österlund, L., Zhdanov, V. P. & Kasemo, B. Photoinduced desorption of potassium atoms from a two-dimensional overlayer on graphite. *J. Chem. Phys.* **106**, 982–1002 (1997).
22. Wilde, M., Beauport, L., Al-Shamery, K. & Freund, H.-J. Photoinduced processes on alkali covered surface: NO desorption from K/Cr₂O₃ (0001). *Surf. Sci.* **390**, 186–193 (1997).
23. Ageev, V. N., Kuznetsov, Yu. A., Yakshinskiy, B. V., Madey, T. E. Electron stimulated desorption of alkali metal ions and atoms: Local surface field relaxation. *Nucl. Instrum. Meth. Phys. Res. B* **101**, 69–72 (1995).
24. Ageev, V. N., Kuznetsov, Yu. A. & Madey, T. E. Electron-stimulated desorption of sodium atoms from an oxidized molybdenum surface. *Phys. Rev. B* **58**, 2248–2252 (1998).
25. Mendillo, M., Baumgardner, J. & Wilson, J. Observational test for the solar wind sputtering origin of the Moon's extended sodium atmosphere. *Icarus* **137**, 13–23 (1999).
26. Heath, D. F. & Thekaekara, M. P. in *The Solar Output and its Variation* (ed. White, O. R.) 193–212 (Colorado Assoc. Univ. Press, Boulder, 1977).
27. Kivelson, M. G. & Russell, C. T. *Introduction to Space Physics* 92 (Cambridge Univ. Press, 1995).
28. Zhou, X. L., Zhu, X. Y. & White, J. M. Photochemistry at adsorbate/metal interface. *Surf. Sci. Rep.* **13**, 73–220 (1991).
29. Verami, S., Barbieri, C., Benn, C. & Cremonese, G. Possible detection of meteor stream effects on the lunar sodium atmosphere. *Planet. Space Sci.* **46**, 1003–1006 (1998).

Acknowledgements. We thank R. E. Johnson, A. L. Sprague, D. M. Hunten, J. Kolodziej and V. N. Ageev for discussion and correspondence. This work was supported in part by the Planetary Atmospheres Division of NASA.

Correspondence and requests for materials should be addressed to T.E.M. (e-mail: madey@physics.rutgers.edu).

Systematic enumeration of crystalline networks

Olaf Delgado Friedrichs*, Andreas W. M. Dress*†, Daniel H. Huson‡, Jacek Klinowski§ & Alan L. Mackay||

* FSP Mathematisierung-Strukturbildungsprozesse, Universität Bielefeld, D-33501 Bielefeld, Germany

† Department of Chemical Engineering, City College, CUNY, Convent Avenue at 140th Street, New York, New York 10031, USA

‡ Department of Mathematics, Princeton University, Princeton, New Jersey 08544-1000, USA

§ Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

|| Department of Crystallography, Birkbeck College (University of London), Malet Street, London WC1E 7HX, UK

The systematic enumeration of all possible networks of atoms in inorganic structures is of considerable interest. Of particular importance are the 4-connected networks (those in which each atom is connected to exactly four neighbours), which are relevant to a wide range of systems—crystalline elements, hydrates, covalently bonded crystals, silicates and many synthetic compounds. Systematic enumeration is especially desirable in the study of zeolites and related materials, of which there are now 121 recognized structural types¹, with several new types being identi-

fied every year. But as the number of possible 4-connected three-dimensional networks is infinite, and as there exists no systematic procedure for their derivation, the prediction of new structural types has hitherto relied on empirical methods (see, for example, refs 2–4). Here we report a partial solution to this problem, based on recent advances in mathematical tiling theory^{5–8}. We establish that there are exactly 9, 117 and 926 topological types of, respectively, 4-connected uninodal, binodal and trinodal networks, derived from simple tilings based on tetrahedra. (Here nodality refers to the number of topologically distinct vertices from which the network is composed.) We also show that there are at least 145 more distinct uninodal networks based on a more complex tiling unit. Of the total number of networks that we have derived, only two contain neither three- nor four-membered rings, and most of the binodal and trinodal networks are new.

We define a tiling as a periodic subdivision of space into bounded, connected regions without holes, which we call tiles. If two tiles meet along a surface, we call the surface a face. If three or more faces meet along a curve, we call the curve an edge. Finally, if at least three edges meet at a point, we call that point a vertex. Vertices and edges together form a network. A network can be derived from a tiling whenever it is possible to choose a collection of simple cycles (closed circuits) such that each edge occurs in at least three of them, and the cycles can be spanned by non-intersecting faces such that the union of all faces separates space into tiles. We note the distinction between faces and rings (usually defined as cycles without short cuts⁹).

We have earlier solved the much simpler problem of classifying all periodic tilings of the euclidean plane, the sphere and the hyperbolic plane^{7,10}. Our algorithms¹⁰ enumerate, and permit the visualization of, all possible topological types of tilings for each two-dimensional symmetry group with 1, 2, 3 (and so on) kinds of inequivalent vertices. We shall call these uninodal, binodal, trinodal (and so on) tilings. This approach has been applied to the enumeration of polyoxometalate cages¹¹. We can now address the three-dimensional case, which has direct applications to structural chemistry.

The starting point is to associate with each type of periodic tiling

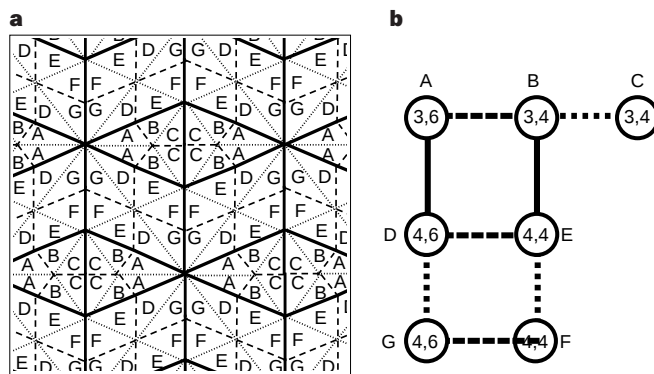


Figure 1 Derivation of the Delaney symbol for a tiling. **a**, The periodic tiling (bold lines) is barycentrically subdivided into triangles. Dotted lines join vertices to tile centres, and dashed lines join edge centres to tile centres. There are seven different classes of triangles, labelled A–G. **b**, Each class of triangle in **a** is represented by a node of the graph. The nodes are joined by lines corresponding to the edges of the triangles. For example, the nodes representing D and E are joined by a dashed line because the corresponding triangles have a dashed edge in common. In addition, each node is labelled by two numbers, the first indicating the number of edges of an original tile containing a triangle of the given class, the second indicating the degree of an original vertex belonging to a triangle of the given class (that is, the number of original edges meeting at this vertex). For example, node G is labelled ‘4,6’ because a triangle of class G is contained in a four-sided tile and its original vertex is of degree 6. The designation A–G of the triangles is arbitrary, and the final Delaney symbol is unique for each kind of periodic tiling.

a unique ‘Delaney symbol’^{6,7}. This is arrived at by breaking the tiling down into simplices using a barycentric subdivision. (Barycentric subdivision of a two-dimensional tiling is obtained by connecting the face centres to the vertices and edge centres: see Fig. 1a.) Any $n + 1$ points in n -space which do not lie in an $(n - 1)$ dimensional space are the vertices of an n -dimensional simplex. A simplex in two dimensions is thus a triangle, and in three dimensions a tetrahedron. In the two-dimensional case, the tiling is broken down into triangles (Fig. 1a), and each class of symmetrically equivalent triangles is then represented by a vertex of a graph. Whenever two triangles are adjacent, the corresponding vertices are joined by an edge. The Delaney symbol is obtained from the graph by appropriate labelling of vertices and ‘colouring’ of edges (Fig. 1b), and can alternatively be written as a string of characters, an ‘inorganic gene’. The ‘gene’ for the tiling in Fig. 1 is

BAD36 ACE34 CBC34 EGA46 DFB44 GEF44 FDG46

where each group of characters contains information about one class of triangles in the barycentric subdivision: group 1 for class A, group 2 for class B, and so on. The letters refer to the neighbours across dashed, dotted and solid lines, respectively, of this class of triangles, while the numbers are the same as those inside circles in Fig. 1b.

The criteria by virtue of which the Delaney symbol encodes a tiling are as follows. (1) A given tiling translates to a linear sequence of characters, which can be put into a unique order by a rule, and vice versa. (2) This sequence is unique to the tiling. (3) There are rules to decide whether a sequence is ‘legal’, that is, corresponds to some actual tiling (see Supplementary Information). The classification of all periodic tilings of a given kind then reduces to the enumeration of the corresponding Delaney symbols, which is equivalent to ‘mutating’ the ‘inorganic gene’. In principle, this approach can be used in any number of dimensions by dividing polytopal tiles into simplices.

In two dimensions, there are 11 different topological types of

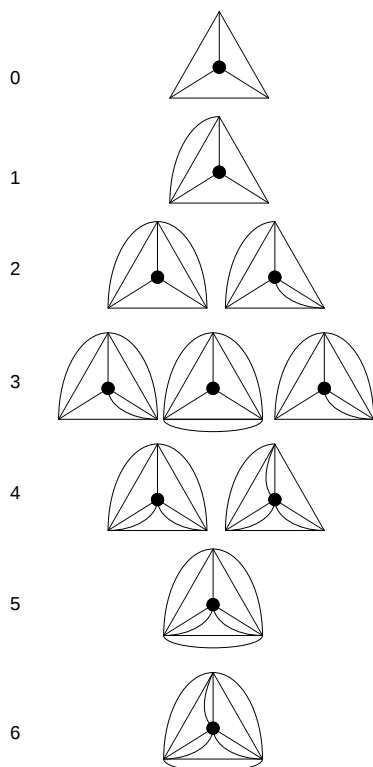


Figure 2 Vertex figures. The tetrahedral vertex figure with no extra edges (top) gives rise to ‘simple’ tilings; vertex figures with extra edges give ‘quasi-simple’ tilings. The number of extra edges is indicated on the left.

uninodal tilings, a result already known to Kepler¹², and 508 and 16,774 types of binodal and trinodal tilings, respectively¹⁰. By contrast, the number of types of uninodal tilings in three dimensions is infinite¹³. However, the number of possibilities becomes finite if we also limit the number of faces which can be incident to a common vertex or, alternatively, the number of faces which can be incident to a common tile (as well as the number of symmetrically inequivalent tiles). In the second case, the set of boundary faces of a given tile forms a two-dimensional tiling, which can be described by a two-dimensional Delaney symbol. The configuration of edges, faces and tiles around a fixed vertex can also be described by a two-dimensional Delaney symbol via the so-called vertex figure: place the centre of a small notational sphere at the vertex and consider the tiling of that sphere formed by the intersections with the different tiles touching that vertex. Some possible vertex figures for 4-connected networks are shown in Fig. 2.

A specific vertex figure can be described in terms of a two-dimensional Delaney symbol (because it is a tiling of the sphere). To enumerate all possible tilings based on a given vertex figure, we must consider all possible ‘three-dimensional’ extensions of its two-dimensional symbol. The main difficulty is to determine which of the resulting three-dimensional symbols give rise to euclidean tilings (as opposed to tilings of, say, hyperbolic or spherical space). A practical solution has been found¹⁴, by combining methods and results from group theory, topology and combinatorics.

Delaney symbols translate classification of tilings into purely algebraic problems. It only remains to develop and implement algorithms which solve them. Our computer program considers all possible ‘legal’ permutations of the ‘gene’, and then generates the corresponding tiling. We first enumerate all possible vertex figures of the desired kind, using our two-dimensional methods. Second, using advanced branch-and-bound techniques and the properties of crystallographic groups, we generate all possible three-dimensional extensions by generating all admissible face pairings and edge degrees for each feasible combination of vertex figures. Third, using methods from combinatorial group theory and topology, we determine whether such a tiling can be realized in three-dimensional euclidean space. Finally, if this is the case, we use optimization methods to construct a plausible realization of the actual tiling (see Supplementary Information).

Using this approach, we have been able to give a complete enumeration of all topological types of simple uninodal, binodal and trinodal tilings. The various networks generated in this way were compared with known structures using ‘coordination sequence’ fingerprinting^{15,16}. Thus in a 4-connected network, each vertex is connected to $N_1 = 4$ neighbouring vertices. These are linked to N_2 vertices in the next shell, in turn connected to N_3 vertices and so on, including each vertex only once. Although the coordination sequence for each kind of vertex is not unique to a given structure, it is a useful guide as structures with different coordination sequences must be different.

Simple tilings can give rise only to structures composed entirely of cages. We have considered those first, especially since networks with the lowest density tend to derive from simple tilings. We have established that there are nine possible topological types of simple uninodal tilings of euclidean space, each corresponding to a 4-connected network^{17,18}. Six of them correspond to known zeolites (structure types SOD, LTA, RHO, FAU, KFI and CHA) and the remaining three have been described by O’Keeffe³. A systematic enumeration of all quasi-simple uninodal tilings yields 285 additional topological types. The total list of 294 tilings gives rise to at least 154 different 4-connected networks, among them the 12 remaining uninodal zeolites.

The enumeration of structure types with two or more symmetrically inequivalent types of tetrahedral vertices (binodal, trinodal, and so on) is addressed in a similar manner. We have established that there are exactly 117 and 926 topological types of simple

binodal and trinodal tilings of euclidean space, respectively. Some examples are shown in Fig. 3. Only two, corresponding to structure types MEP (melanophlogite) and MTN (dodecasil-3C and zeolite ZSM-39), both of them silica polymorphs, contain neither 3- nor 4-membered rings. Whereas most of the uninodal networks are known^{3,4,19}, most of the others are new.

Known zeolitic structure types involve n -nodal structures with n up to 12, found in zeolite ZSM-5 (MFI structure type), one of the most complex inorganic structures known. Given the vast number of possible combinations of various vertex figures, MFI is not expected to be found among the first few million structures enumerated. All the same, the method is systematic and exhaustive, and will eventually deliver the ZSM-5 structure. Further, only a small fraction (possibly a finite number) of the many structures will be chemically feasible (with bond lengths and angles within an acceptable range)²⁰. Given a realizable hypothetical structure, a least-squares fit leading to optimal atomic positions could then be performed by computer. We are investigating criteria for the automatic exclusion of chemically implausible structures at an early stage. One method is based on the detection of energetically unfavourable features. For example, there are 267 quasi-simple triangle-free uninodal tilings and 69 and 301 triangle-free binodal and trinodal simple tilings, respectively.

A further consideration concerns the well known method of constructing larger 4-connected tilings from smaller tilings by replacing certain faces by prisms or stacks of prisms. The simplest case is replacing a square by a stack of cubes. Such tilings can have very large faces and tiles, but no realizations if almost equal edge lengths and admissible angles are simultaneously required. Thus, where prisms meet along their bases, one angle close to 180° cannot be avoided. There are exactly 53 binodal and 230 trinodal triangle-free and prism-stack-free or, for short, 'relaxed' simple tilings. The largest face sizes among those are 12 and 24, while the largest tiles contain 74 and 98 faces, respectively. Of these tilings, there are exactly four binodal tilings and 15 trinodal tilings, in which all tiles are symmetrically equivalent and therefore congruent. Some examples are shown in Fig. 3.

Another issue is how to deal with distinct tilings leading to equivalent networks. We have applied simple techniques to eliminate repetition. With the help of coordination sequences^{15,16} and ring statistics⁹ the list of solutions can usually be reduced to very small sublists, each containing only tilings corresponding to identical networks.

Our results and methods are applicable to the structure of zeolites, silicate networks substituted with heteroatoms (such as Al, Ga and Be), aluminophosphates and related materials, nitrides,

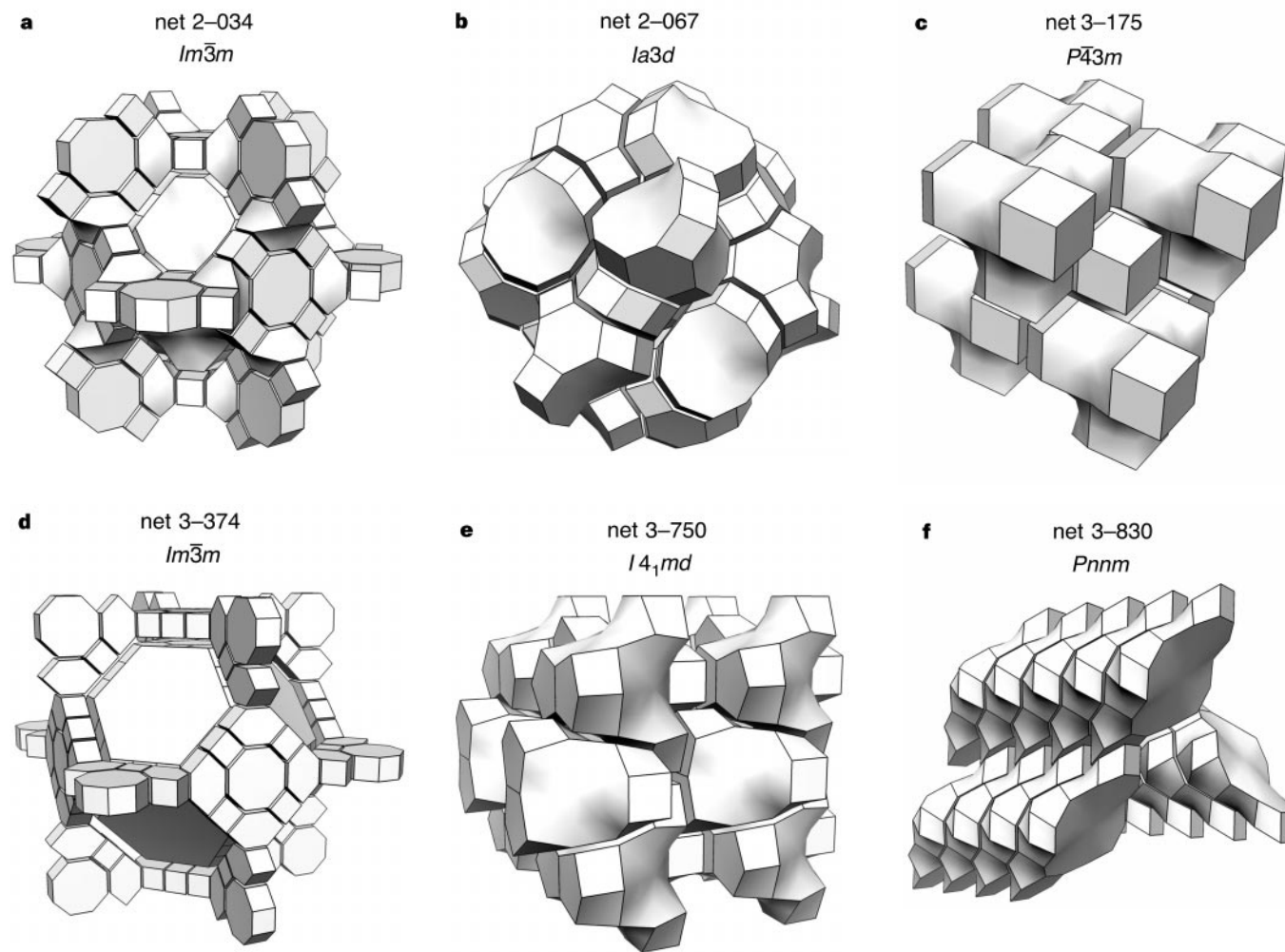


Figure 3 Some of the simple binodal (**a**, **b**) and trinodal (**c**–**f**) structures enumerated, together with their space groups. **a**, Relaxed binodal tiling with the largest tile (74 faces, 144 vertices and 216 edges); **b**, one of four relaxed binodal tilings with exactly one kind of symmetrically inequivalent tile, and the only one containing 10-gons; **c**, the only relaxed trinodal tiling with exactly two types of

symmetrically inequivalent faces; **d**, relaxed trinodal tiling with the largest tile (98 faces, 192 vertices, 288 edges, largest face size 18); **e**, one of nine relaxed trinodal tilings with 9-gons; **f**, the only relaxed trinodal tiling with 14-gons. The pictures were generated using Geomview software (version 1.6.1 (1996); The Geometry Centre, Minnesota, USA).

chalcogenides and halides, to 3- and 4-connected carbon networks and to structures such as ice. Finally, bubbles found in foams are polyhedra in simple tilings. This means that every periodic foam with 1, 2 and 3 kinds of vertex will be found among our total of 1,052 ($= 9 + 117 + 926$) simple tilings. □

Received 18 May; accepted 28 June 1999.

1. Meier, W. M., Olson, D. H. & Baerlocher, C. *Atlas of Zeolite Structure Types* (Elsevier, London, 1996); updates available at (<http://www.iza-sc.ethz.ch/IZA-SC/Atlas/AtlasHome.html>).
2. Smith, J. V. Topochemistry of zeolites and related materials. 1. Topology and geometry. *Chem. Rev.* **88**, 149–182 (1988).
3. O'Keeffe, M. & Hyde, B. G. *Crystal Structures I: Patterns and Symmetry* (Monograph, Mineralogical Assoc. of America, Washington DC, 1996).
4. Treacy, M. M. J., Randall, K. H., Rao, S., Perry, J. A. & Chadi, D. J. Enumeration of periodic tetrahedral frameworks. *Z. Kristallogr.* **212**, 768–791 (1997).
5. Dress, A. W. M., Huson, D. H. & Molnár, E. The classification of face-transitive periodic 3-dimensional tilings. *Acta Crystallogr. A* **49**, 806–817 (1993).
6. Dress, A. W. M. in *Springer Lecture Notes In Mathematics* Vol. 1172, 56–72 (Springer, Göttingen, 1985).
7. Dress, A. W. M. Presentations of discrete groups, acting on simply connected manifolds, in terms of parametrized systems of Coxeter matrices: a systematic approach. *Adv. Math.* **63**, 196–212 (1987).
8. Huson, D. H. Combinatorial tiling theory: theorems-algorithms-visualization. At (<http://www.mathematik.uni-bielefeld.de/~delgado/Tiling-Theory>).
9. O'Keeffe, M. & Brese, N. E. Uninodal 4-connected 3D nets. I. Nets without 3-rings or 4-rings. *Acta Crystallogr. A* **48**, 663–669 (1992).
10. Huson, D. H. The generation and classification of tile- k -transitive tilings of the Euclidean plane, the sphere and the hyperbolic plane. *Geometriae Dedicata* **47**, 269–296 (1993).
11. Delgado Friedrichs, O., Dress, A. W. M., Müller, A. & Pope, M. T. Polyoxometalates: a class of compounds with remarkable topology. *Mol. Eng.* **3**, 9–28 (1993).
12. Kepler, J. *Harmonices Mundi Libri V* Book II, Sections 18–20, pp 51–56 (Linz, 1619).
13. Heesch, H. Über Raumteilungen. *Nachr. Gesellschaft Wiss. Göttingen* **1**, 35–42 (1934).
14. Delgado Friedrichs, O. Euclidicity criteria for three-dimensional branched triangulations. Thesis, Univ. Bielefeld (1994).
15. Brunner, G. O. & Laves, F. Zum Problem der Koordinationszahl. *Wiss. Z. Tech. Univ. Dresden* **20**, 387–390 (1971).
16. Meier, W. M. & Moock, H. J. The topology of three-dimensional 4-connected nets: classification of zeolite framework types using coordination sequences. *J. Solid State Chem.* **27**, 349–355 (1979).
17. Wells, A. F. *Three-Dimensional Nets and Polyhedra* (Wiley, New York, 1977).
18. Wells, A. F. *Further Studies of Three-Dimensional Nets* (Am. Crystallographic Assoc. Monogr. No. 8, Polycrystal Book Service, Pittsburgh, 1979).
19. Grünbaum, B. Uniform tilings of 3-space. *Geombinatorics* **4**, 49–56 (1994).
20. Gramlich-Meier, R. & Meier, W. M. Constituent units and framework conformations in zeolite networks. *J. Solid State Chem.* **44**, 41–49 (1982).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank M. O'Keeffe for discussions.

Correspondence and requests for materials should be addressed to J.K. (e-mail: jk18@cam.ac.uk).

The mechanisms for pressure-induced amorphization of ice I_h

J. S. Tse*, D. D. Klug*, C. A. Tulk*, I. Swainson*,
E. C. Svensson*, C.-K. Loong†, V. Shpakov‡,
V. R. Belosludov‡, R. V. Belosludov§ & Y. Kawazoe§

* Steacie Institute for Molecular Sciences, National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6

† Argonne National Laboratory, Argonne, Illinois 60439, USA

‡ Inorganic Chemistry Institute, Russian Academy of Sciences, Novosibirsk, Russia

§ Institute for Materials Research, Tokoku University, Sendai 980-8577, Japan

There has been considerable interest in the structure of liquid water at low temperatures and high pressure following the discovery of the high-density amorphous (HDA) phase of ice I_h (ref. 1). HDA ice forms at a pressure close to the extrapolated melting curve of ice, leading to the suggestion that it may have structure similar to that of dense water. On annealing, HDA ice transforms into a low-density amorphous (LDA) phase with a distinct phase boundary^{2,3}. Extrapolation of thermodynamic data along the HDA–LDA coexistence line into the liquid region has led to the hypothesis that there might exist a second critical point for water and the speculation that liquid water is mixture of two distinct structures with different densities^{4,5}. Here we critically examine this hypothesis. We use quasi-harmonic lattice-dynamics

calculations to show that the amorphization mechanism in ice I_h changes from thermodynamic melting for $T > 162$ K to mechanical melting at lower temperatures. The vibrational spectra of ice I_h , LDA ice and quenched water also indicate a structure for LDA ice that differs from that of the liquid. These results call into question the validity of there being a thermodynamic connection between the amorphous and liquid phases of water.

We used the quasi-harmonic lattice-dynamics⁶ method to calculate the thermodynamic and mechanical stability boundaries of ice I_h under pressure. A thermodynamic solid–liquid stability boundary is determined by the equality of free energies of the two phases. In contrast, the mechanical stability boundary is obtained from purely mechanical stability conditions. To test for mechanical stability of the crystal, we calculated elastic constants at several pressures and temperatures. The maximum pressure of stability at a given temperature is determined via the Born stability criteria⁷. A mechanical instability in a solid occurs when particular combinations of the elastic constants violate one of the Born stability conditions⁷. Here we use a semi-empirical approach, combining the quasi-harmonic lattice dynamics and the Lindemann criterion for melting, to compute the thermodynamic melting line of ice I_h . This approach⁸ has been applied successfully to a number of systems, including the correct description of the water–ice VI phase boundary. (For a recent application on related systems see refs 9, 10.)

The mechanical instability line due to the violation of the Born stability condition $C_{11} - |C_{12}| > 0$ (ref. 7) and the theoretical thermodynamic curves are compared with experiment in Fig. 1. The calculated thermodynamic melting curve agrees well with experiment¹¹ in the low-pressure/high-temperature region; the mechanical instability line agrees well with experiment in the high-pressure/low-temperature region. The temperature where mechanical instability occurs is always higher than the thermodynamic melting point. The most significant theoretical result is that the thermodynamic melting line meets the mechanical instability curve at ~ 160 K. This result, which agrees with experiment, shows that there are two distinct mechanisms for pressure-induced transformation of ice I_h in two different temperature regimes. This behaviour can be rationalized as follows. At high temperature, the water molecules have sufficiently large-amplitude thermal motions and consequently, the thermodynamic melting process takes precedence over mechanical instability. In contrast, at temperatures lower than 160 K the vibrational amplitudes of the water molecules are reduced, and the crystal structure collapses due to a mechanical instability under pressure.

In clathrate hydrates and their zeolite silica analogues, the mechanism for transformations to high-density amorphous phases at low temperatures has also been established to be a mechanical instability^{12,13}. In contrast to ice I_h , the high-density phase made after pressurization in the clathrate hydrates of tetrahydrofuran (THF) and sulphur hexafluoride¹² could not be recovered even at 77 K. Instead, the transformed phase immediately reverted back to the initial crystalline solid. This reversible transformation would certainly not be feasible if the clathrate had actually undergone a melting transition to a quenched solution, as the formation kinetics of clathrate hydrate at 77 K are extremely slow¹². This mechanism has been clearly demonstrated in a combined theoretical and experimental study of pressure-induced transformations in a THF zeolite¹³.

Thermodynamic analysis of the extrapolated melting curve for ice I_h has shown¹⁴ that the experimental results at high temperature and low pressure can be reproduced using reasonable parameters. However, this approach fails in the high-pressure/low-temperature region. For ice I_h the predicted amorphization pressure at 80 K of 6 kbar is substantially lower than the 11 kbar observed in experiments. A mechanical instability due to the softening of the elastic modulus C_{66} ($= C_{11} - C_{12}$) of the ice structure under high pressure has been proposed in a molecular dynamics study¹⁵.

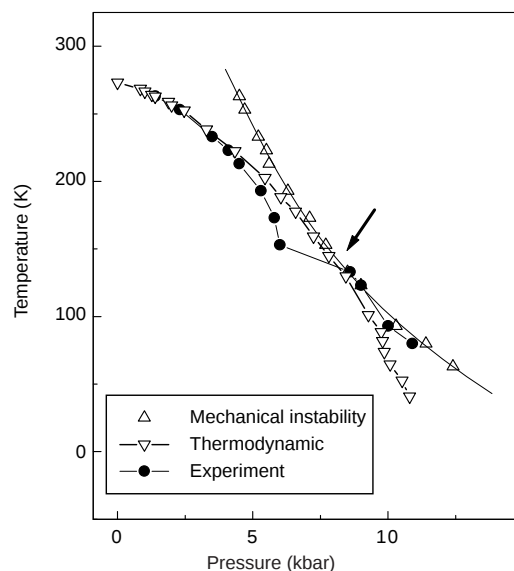


Figure 1 Phase behaviour (experimental and theoretical) of ice I_h . Here we compare the experimental results on the transition pressure with the theoretical calculated thermodynamic melting line and mechanical instability line of ice I_h . Lattice dynamics calculations were performed on a zero dipole moment simulation cell consisting of 64 water molecules employing the TIP4P inter-molecular potential²³ with extensive Brillouin zone $\mathbf{k}$ -point sampling to ensure convergence in the calculated quantities. The temperature and pressure dependencies of the elastic constants are calculated with the method adapted recently for molecular crystals using the approximation of rigid molecules⁶. The thermal expansion and elastic constants are determined from the appropriate derivatives to the free energy F_{th} of the crystal, $F_{\text{th}} = U + F_s$, where U is the potential energy and F_s is the vibrational part of the free energy. In particular, the second derivatives of the potential energy with respect to strain can be described by the elements of the dynamical matrix of the crystal and its derivatives of the wavevector $\mathbf{q}$ at the $\mathbf{q} = 0$ point. The approximation of the neglect of the purely anharmonic term does not greatly affect the resulting calculation of elastic constants, as has been demonstrated in recent studies²⁴. The melting point of a crystal can be determined from a consideration of the maximum root-mean-square amplitudes of molecular vibrations. A scaling factor of 0.16 to match the theoretical melting point to the experimental melting point, used in subsequent calculations, was determined empirically by matching the experimental melting point to the calculated ice I_h vibrational amplitude at 273 K. The arrow indicates the cross-over of the two mechanisms.

The transformation to a dense phase is due to the compaction of the ice structure via the filling of the interstitial voids, and does not correspond to the formation of a liquid-like structure. The fundamental assumption that water and the amorphous ice phases have a close thermodynamic connection implies that the structure of liquid water can be described at any given temperature and pressure by a mixture of the HDA and LDA ice structures¹⁶. The assumption is based on speculation that the neutron structure factors of HDA and LDA ice are similar to liquid water. Careful inspection of the structure factors of liquid water indicates that the short-range region is often poorly described by a weighted sum of the HDA and LDA ice. In a few cases, unrealistic weighting coefficients greater than unity have to be used in order to obtain agreement with the liquid-water structure¹⁶.

The fact that the mechanism of the transformation of ice I_h to HDA ice has been established to be due to a mechanical instability casts doubt on the assumption that HDA ice is thermodynamically connected to a liquid state. As there is a distinct phase boundary³ between HDA and LDA ice, the suggestion that amorphous ices are connected to liquid water lead one to expect that the structures of

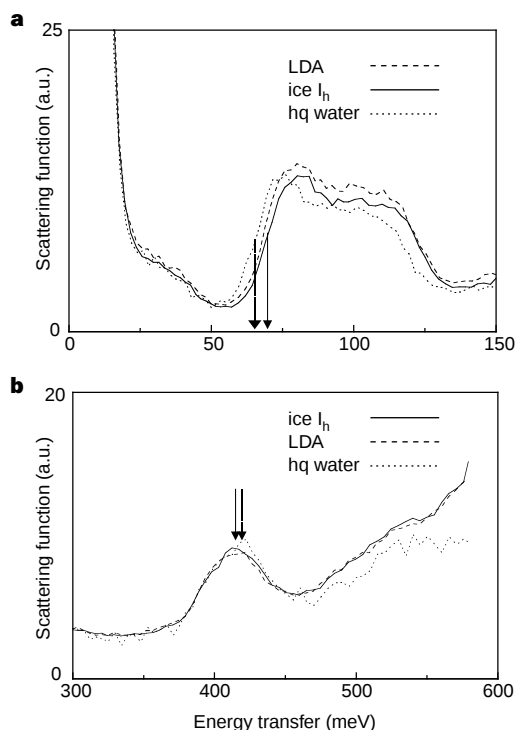


Figure 2 The experimental inelastic incoherent neutron scattering (IINS) function for LDA ice, hyperquenched (hq) water, and ice I_h . Data are shown in librational (**a**) and internal vibration (**b**) regions, illustrating the different spectral responses of these materials. The arrows indicate the onset energies in **a**, and the mean frequencies in **b**. The error limits in the IINS measurements are ± 1.1 meV in the vibrational frequency and ± 0.2 in the intensity units. The quenched water was prepared by rapid cooling (10^6 deg s^{-1}) of 3- μm -diameter water droplets on a cold Cu substrate, as described previously²², which gives a glassy form of water that can be studied at low temperatures. The rapid cooling avoids the crystallization that would occur at lower cooling rates. Details of the inelastic scattering data will be published elsewhere. a.u., arbitrary units.

liquid water (at the same densities as the amorphous ices) quenched below the glass transition points should share common structural features with HDA and LDA ices. X-ray and neutron diffraction have been performed on both HDA and LDA ice^{17–20}, and the results can be compared with published data for high-density water²¹ and rapidly quenched liquid water²⁰. At 7.7 kbar, the density of liquid water is close to that of HDA ice studies at 77 K (ref. 21). The first experimental claim that the structure of HDA ice is similar to liquid water was made by Floriano *et al.*¹⁷. However, careful inspection of the neutron diffraction results revealed discrepancies in the short-range region. Differences in the positions and relative amplitudes of the peaks in the range 3–7 Å in the O–O pair correlation function between HDA ice^{18,19} and dense liquid water²¹ were also reported in more recent measurements. More specifically, beyond the first nearest neighbour distance, a peak at 3.2 Å was observed in water at 7.7 kbar, whereas in HDA ice the peak is located at 3.7 Å; a peak at 4.25 Å in liquid water is in contrast to the peak at 4.65 Å in HDA ice. Another remaining question is whether LDA ice is structurally related to the quenched liquid. Diffraction results show subtle differences between LDA and rapidly quenched water. The main differences²⁰ between LDA ice and water quenched rapidly at 77 K are that the first peak in LDA ice in the O–O pair correlation function is higher for LDA ice, and that there is a minimum located at ~ 5.8 Å. These differences are significant as they show that the local structural organization in the interstitial region is different. Since the diffraction method is a probe for long-range correlation, a

more definitive characterization of the short-range order requires a technique sensitive to the local structure. The short-range order and intermolecular interactions in hydrogen-bonded systems can be obtained from the vibrational spectra of these systems in the energy range of internal vibrations of water molecules.

To investigate the short-range order of amorphous ice the vibrational spectra were obtained from inelastic incoherent neutron scattering (IINS) experiments. This method provides a very sensitive measurement of the near-neighbour interactions (hydrogen-bond strengths). In Fig. 2 we show the IINS scattering function at 10 K for LDA ice, quenched water and ice I_h for libration (Fig. 2a) and internal modes (Fig. 2b) of water molecules. The librational vibration bands in the energy range 55–135 meV are determined primarily by the intermolecular hydrogen bonding that restricts rotation of individual molecules (Fig. 2a). The scattering functions are characterized by a rather sharp onset and a dominant peak followed by a broad featureless band. For LDA ice and ice I_h , the respective onset energy and position of the first dominant peak, as well as the full width at half maximum (FWHM), are nearly identical. The band profile of quenched water is distinctively different from that of LDA ice and ice I_h : the FWHM is less (52 meV versus 60 meV), with the low-energy edge shifted by 4 meV to lower frequency and the high-energy cut-off decreased by 6 meV. The lower onset frequency indicates that the average hydrogen-bond strength is less in quenched water than in LDA ice and ice I_h and, therefore, the local structure also differs.

The scattering function in the internal-mode region reaffirms the trend of hydrogen-bond strength in LDA ice, quenched water and ice I_h . For the intramolecular (internal) vibrations in the energy range 300–600 meV, the scattering functions are characterized by a single broad and nearly gaussian band. The large bandwidth is the result of strong inter- and intra-molecular couplings between internal vibrations. The vibrational profile for LDA ice is again almost identical to that of ice I_h but differs from that of quenched water (Fig. 2b). The most distinctive difference is that the frequency distribution in quenched water is shifted to higher frequency by ~5 meV as compared with LDA ice and ice I_h . The higher frequency shift indicates weaker hydrogen bonding and implies different short-range order in quenched water from that in LDA ice and ice I_h . On the other hand, the similarity in the scattering function of ice I_h and LDA ice shows that both materials have similar hydrogen-bonding interactions with their nearest neighbours. This assertion is validated by a recent Raman-scattering study²² on uncoupled O–H or O–D oscillators in dilute solutions of H_2O in D_2O or D_2O in H_2O , in quenched water, LDA ice and ice I_h . In contrast to IINS, Raman spectroscopy measures with very high sensitivity the frequency of a particular localized oscillator which is determined primarily by the near-neighbour O–H...O interaction. For LDA ice, ice I_h and quenched water measured at 15 K, the Raman frequencies for the O–H/O–D vibrations are 406.7/300.7 meV, 404.7/299.2 meV, and 409.4/302.5 meV, respectively. The frequencies of the uncoupled O–H or O–D modes in LDA ice are quite different from quenched water, indicating that their local hydrogen-bond interactions are not the same.

The evidence presented here raises questions concerning a thermodynamic connection between amorphous ices and liquid water. Our results indicate that the pressure-induced phase transformation of ice I_h at low temperatures is not a melting transition in the thermodynamic sense, and HDA therefore is not expected to have a liquid-like structure. In addition, the vibrational spectra obtained from Raman spectroscopy and IINS show that there are distinct differences between LDA ice and quenched liquid water. Future testing of thermodynamic connections between the pressure-induced amorphous ice and the high-density liquid could certainly include other methods such as reverse Monte Carlo analysis of both neutron and X-ray data for the various amorphous-ice structures. □

Received 27 January; accepted 18 June 1999.

- Mishima, O., Calvert, L. D. & Whalley, E. 'Melting' ice I at 77 K and 10 kbar: a new method of making amorphous solids. *Nature* **310**, 393–395 (1984).
- Mishima, O., Calvert, L. D. & Whalley, E. An apparently first-order transition between two amorphous phases of ice induced by pressure. *Nature* **314**, 76–78 (1985).
- Mishima, O. Reversible first-order transition between the two H_2O amorphs at 0.2 GPa and 135 K. *J. Chem. Phys.* **100**, 5910–5912 (1994).
- Mishima, O. & Stanley, E. H. Decompression-induced melting of ice IV and the liquid–liquid transition in water. *Nature* **392**, 164–168 (1998).
- Mishima, O. & Stanley, E. H. The relationship between liquid, supercooled and glassy water. *Nature* **396**, 329–335 (1992).
- Shpakov, V. P., Tse, J. S., Belosludov, V. R. & Belosludov, R. V. Elastic moduli & instability in molecular crystals. *J. Phys. Condens. Matter* **9**, 5853–5865 (1997).
- Born, M. & Huang, K. *Dynamical Theory of Crystal Lattices* (Oxford Univ. Press, London, 1954).
- Bruesch, P. *Phonons: Theory and Experiments I* (Springer, New York, 1982).
- Belosludov, R. V., Grochev, E. V., Dyanin, Yu. A. & Belosludov, V. R. in *Proc. 2nd Int. Conf. on Natural Gas Hydrate* 303–309 (PROGE, Toulouse, 1996).
- Tse, J. S., Shpakov, V. P. & Belosludov, V. R. High-pressure elastic constants of solid krypton from quasi-harmonic lattice-dynamics calculations. *Phys. Rev. B* **58**, 2365–2368 (1998).
- Mishima, O. Relationship between melting and amorphization of ice. *Nature* **384**, 546–549 (1996).
- Handa, Y. P., Tse, J. S., Klug, D. D. & Whalley, E. Pressure-induced phase transitions in clathrate hydrates. *J. Chem. Phys.* **94**, 623–627 (1991).
- Tse, J. S., Klug, D. D., Ripmeester, J. A., Desgreniers, S. & Lagarec, K. The role of non-deformable units in pressure-induced reversible amorphization of clathrals. *Nature* **369**, 724–727 (1994).
- Whalley, E., Klug, D. D. & Handa, Y. P. Entropy of amorphous ice. *Nature* **342**, 782–783 (1989).
- Tse, J. S. Mechanical instability in ice I_h : A mechanism for pressure-induced amorphization. *J. Chem. Phys.* **96**, 5482–5487 (1992).
- Bellissent-Funel, M. C. Is there a liquid–liquid phase transition in supercooled water? *Europhys. Lett.* **42**, 161–166 (1998).
- Floriano, M. A., Whalley, E., Svensson, E. C. & Sears, V. F. Structure of high-density amorphous ice by neutron diffraction. *Phys. Rev. Lett.* **57**, 3062–3064 (1986).
- Bosio, L., Johari, G. P. & Teixeira, J. X-ray study of high-density amorphous water. *Phys. Rev. Lett.* **56**, 460–463 (1986).
- Bizid, A., Bosio, L., Defrain, A. & Oumezzine, M. Structure of a high-density amorphous water. I. X-ray diffraction. *J. Chem. Phys.* **87**, 2225–2230 (1987).
- Bellissent-Funel, M. C., Bosio, L., Hallbrucker, A., Mayer, E. & Sridi-Dorbez, R. X-ray and neutron scattering studies of the structure of hyperquenched glassy water. *J. Chem. Phys.* **97**, 1282–1286 (1992).
- Okhulkov, A. V., Demianets, Y. N. & Gorbaty, Y. E. X-ray scattering in liquid water at pressures of up to 7.7 kbar: Test of a fluctuation model. *J. Chem. Phys.* **100**, 1578–1588 (1994).
- Tulk, C. A., Klug, D. D., Branderhorst, R., Sharpe, P. & Ripmeester, J. A. Hydrogen bonding in glassy liquid water from Raman spectroscopic studies. *J. Chem. Phys.* **109**, 8478–8484 (1998).
- Jorgensen, W. L., Chandrasekhar, J., Madura, R. W., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–936 (1983).
- Sprick, M., Impey, R. I. & Klein, M. L. Second order elastic constants for Lennard-Jones solid. *Phys. Rev. B* **29**, 4368–4674 (1984).

Acknowledgements. Inelastic scattering data were acquired at Argonne National Laboratory: these measurements were supported by the US DOE-BES.

Correspondence and requests for materials should be addressed to J.S.T. (e-mail: John.Tse@NRC.CA)

A direct-methane fuel cell with a ceria-based anode

E. Perry Murray, T. Tsai & S. A. Barnett

Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA

Fuel cells constitute an attractive power-generation technology that converts chemical energy directly and with high efficiency into electricity while causing little pollution. Most fuel cells require hydrogen as the fuel, but viable near-term applications will need to use the more readily available hydrocarbons, such as methane. Present-day demonstration power plants and planned fuel-cell electric vehicles therefore include a reformer that converts hydrocarbon fuel into hydrogen. Operating fuel cells directly on hydrocarbons would obviously eliminate the need for such a reformer and improve efficiency. In the case of polymer-electrolyte fuel cells, which have been studied for vehicle applications, the direct use of methanol fuel has been reported, but resulted in fuel permeating the electrolyte^{1,2}. Solid oxide fuel cells—promising candidates for stationary power generation—can also use hydrocarbon fuel directly to generate energy, but this mode of operation resulted in either carbon deposition at high temperatures or poor power output at low operating temperatures^{3–5}. Here we report the direct electrochemical oxidation of methane in solid oxide fuel

cells that generate power densities up to 0.37 W cm^{-2} at 650°C . This performance is comparable to that of fuel cells using hydrogen^{6,7} and is achieved by using ceria-containing anodes and low operating temperatures to avoid carbon deposition. We expect that the incorporation of more advanced cathodes would further improve the performance of our cells, making this solid oxide fuel cell a promising candidate for practical and efficient fuel-cell applications.

Earlier studies of solid oxide fuel cells (SOFCs) have investigated the potential for both the direct use of hydrocarbons and internal reforming (conversion of the hydrocarbon to hydrogen at the SOFC anode). Internal reforming is generally achieved by using nickel-based anodes to catalyse the endothermic reforming reactions. In the case of methane, these are:



where ΔH_{298}° is the standard enthalpy of the reaction at 298 K. However, the challenges in maintaining appropriate gas composition and temperature gradients across large-area SOFC stacks make it difficult to achieve 100% internal reforming^{8,9}. Instead, a

partial external pre-reforming stage is typically used, followed by internal reforming within the stack¹⁰. The direct electrochemical oxidation of hydrocarbon fuels in SOFCs is, in principle, possible, but has resulted in carbon deposition^{3,5} at operating temperatures above about 800°C , and low power densities^{3,4} ($\sim 10 \text{ mW cm}^{-2}$) at temperatures below 800°C . But recently developed SOFCs produce high power densities by using hydrogen fuel at $600\text{--}800^\circ\text{C}$ (refs 6, 7, 11–15) and may thus have the potential to provide good power densities by directly using hydrocarbons in a temperature range where carbon deposition can be avoided.

The SOFCs we report here were fabricated on porous $\text{La}_{0.8}\text{Sr}_{0.2}\text{MnO}_3$ (LSM) cathodes. The LSM pellets were $\sim 2 \text{ cm}$ in diameter and 1 mm thick, and were produced using standard ceramic processing techniques. All SOFC layers, starting with a $0.5\text{-}\mu\text{m}$ -thick $(\text{Y}_2\text{O}_3)_{0.15}(\text{CeO}_2)_{0.85}$ (YDC) porous film, were deposited on the LSM pellet using d.c. reactive magnetron sputtering. The electrolyte, $8 \text{ mol}\%$ Y_2O_3 -stabilized ZrO_3 (YSZ), was then deposited under conditions yielding a dense, $8\text{-}\mu\text{m}$ -thick film. To complete the cell, another $0.5\text{-}\mu\text{m}$ -thick YDC film was deposited, followed by a porous, $2\text{-}\mu\text{m}$ -thick Ni-YSZ anode. Although sputtering was used in this study, similar cells can be made using other methods^{6,11–13}. The SOFCs and their performance on hydrogen fuel are described in detail elsewhere⁷. Anode reactions were studied using impedance

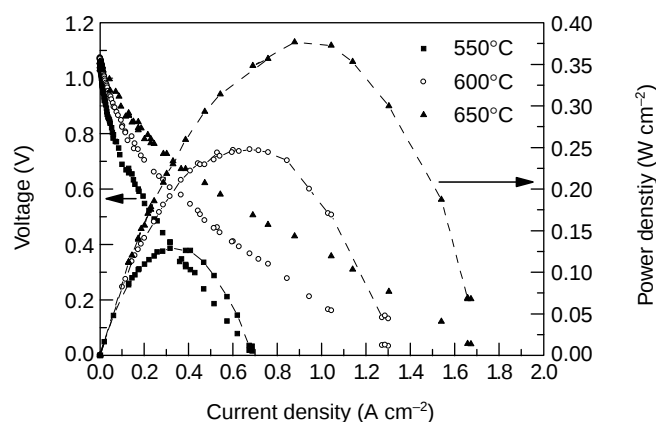


Figure 1 Cell voltage and power density versus current density for an SOFC operated on air and wet methane. The measurements were collected in atmospheric-pressure air, and the methane fuel was supplied at $\sim 50 \text{ cm}^3 \text{ STP min}^{-1}$.

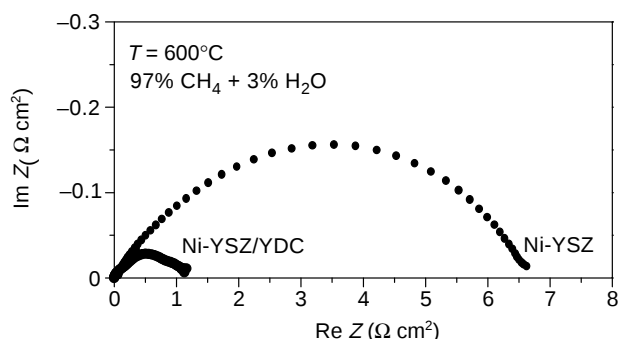


Figure 3 Impedance spectroscopy results for Ni-YSZ and Ni-YSZ/YDC anodes in $97\% \text{CH}_4 + 3\% \text{H}_2\text{O}$ at 600°C . The interfacial resistance corresponds to the difference between the real-axis intercepts.

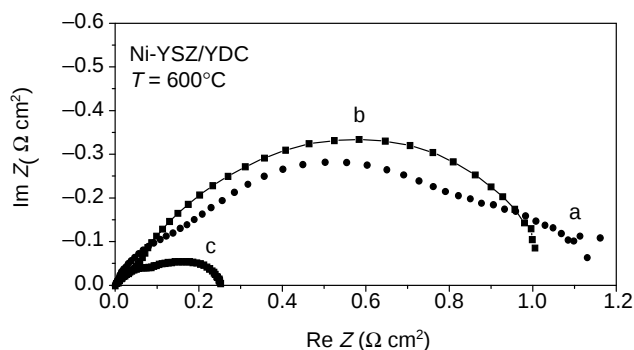


Figure 2 Comparison of electrode impedance spectra for Ni-YSZ/YDC anodes. Spectra were measured at 600°C in $97\% \text{CH}_4 + 3\% \text{H}_2\text{O}$ (trace a), $3\% \text{H}_2 + 3\% \text{H}_2\text{O} + 94\% \text{Ar}$ (trace b) and $97\% \text{H}_2 + 3\% \text{H}_2\text{O}$ (trace c).

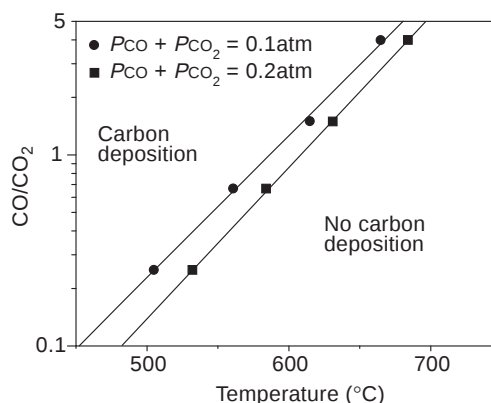


Figure 4 Calculated CO/CO_2 ratios at equilibrium with graphitic carbon, based on reaction (5).

spectroscopy with anodes that were sputter-deposited onto both sides of bulk YSZ single-crystal electrolytes.

Figure 1 shows measurements of current density and power density versus voltage, performed on a typical cell using air and methane. The results for dry and wet (with 3% H₂O) methane were nearly identical. Cell performance was stable in our preliminary 100-h life tests, except for wet methane at low voltages where the anode Ni gradually oxidized. The results in Fig. 1 are similar to those obtained for the same cells operated with humidified hydrogen fuel, except that the power densities are ~20% lower⁷. The previous studies with hydrogen fuel⁷ showed that the current densities were limited primarily by cathode overpotential. Examination of the anodes after the cell tests (by visual observation, energy dispersive X-ray, and scanning electron microscopy) showed no evidence of carbon deposition after ~100 h of operation.

Successful cell operation on dry methane indicated that direct electrochemical oxidation,



was the primary anode reaction. The Gibbs free energy for reaction (3) is given by ΔG° . Methane reforming (reactions (1) and (2)) may have also played a role in the cell operation, but only after H₂O and CO₂ were produced through reaction (3). Even after H₂O and CO₂ production, reforming rates were probably too low to compete with the primary anode reaction, because the small anode area (~1 cm²), low temperature⁸ and flushing of products (H₂O and CO₂) from the anode compartment by the relatively high fuel flow rates suppress reactions (1) and (2).

More evidence of direct methane oxidation is provided by the different impedance spectra (Fig. 2) obtained from these Ni-YSZ/YDC anodes in humidified methane (trace a), humidified dilute H₂ (trace b) and humidified H₂ (trace c). The 3% H₂ + 3% H₂O + 94% Ar mixture used to obtain trace b simulates a slightly reformed methane fuel. This mixture yielded an electrode spectrum with a shape different from that of the methane spectrum (a). Since the shape of the spectra are related to the mechanisms governing the fuel reaction, this result suggests that the primary anode reaction with methane was not oxidation of hydrogen produced by reforming. Trace c shows that the electrode resistance was much smaller for H₂ than for CH₄, explaining why the SOFC current densities were higher for hydrogen than methane.

The rapid direct electrochemical oxidation of methane at these temperatures was due to the anodes employed in these SOFCs¹⁶, which combined Ni-YSZ and YDC layers. Impedance measurements taken in humidified methane for Ni-YSZ/YDC and Ni-YSZ anodes (Fig. 3) indicate that the YDC layer caused the interfacial resistance to reduce by a factor of ~6. This agrees with previous studies that indicate that ceria promotes hydrocarbon oxidation¹⁷. Ceria is beneficial for several reasons. First, it becomes a mixed conductor in the reducing fuel environment¹⁴, a condition which should expand the reaction zone beyond three-phase boundaries. Second, the ionic conductivity of ceria is higher than that of YSZ, which improves the transport of oxygen ions from the electrolyte to the anode. Third, ceria is known to readily store and transfer oxygen, and adding zirconia enhances the storage capability¹⁸. The present anodes have a ceria layer and two ceria/zirconia interfaces where enhanced oxygen storage may increase methane oxidation rates.

An important result from the cell tests was the absence of carbon deposition. In general, carbon deposition can occur by methane pyrolysis,



or carbon monoxide disproportionation,



The role of methane pyrolysis was tested by flowing pure methane over SOFC anodes without operating the SOFC, so that no reaction products were present. No carbon deposition was observed at < 700 °C: that is, there was no methane pyrolysis. The amount of carbon deposited increased with increasing temperature above 700 °C, but less carbon deposition was observed at a given temperature on Ni-YSZ/YDC anodes than on Ni-YSZ.

The present results are for nearly pure methane, as the small-area anodes and relatively high fuel flow rates resulted in small concentrations of reaction products. The situation that would be encountered in a methane SOFC stack is more complicated, as the products of reactions (1)–(3) would be present at substantial partial pressures, allowing carbon deposition by reaction (5). Although tests on SOFC stacks were beyond the scope of this study, we have performed a simple equilibrium calculation to determine the conditions where one would expect carbon-deposition-free stack operation in CO–CO₂ mixtures (Fig. 4). The results show that low CO/CO₂ ratios suppress carbon deposition through reaction (5). The optimal temperature range for SOFC stack operation on dry methane is ~500–700 °C. At temperatures below this range, carbon deposition through reaction (5) will occur unless CO/CO₂ ratios are kept very low. Temperatures ≥ 700 °C would tend to suppress carbon deposition by reaction (5), but would allow carbon deposition by direct pyrolysis of methane (equation (4)). We also note that carbon deposition may be suppressed by oxygen ions reacting with carbon at the anode during SOFC operation⁵. Finally, we point out that some internal reforming would be necessary in a SOFC stack to produce a small amount of CO and H₂. These compounds would balance the CO₂ and H₂O produced by direct oxidation, preventing the too-low CO/CO₂ and H₂/H₂O ratios at which the nickel anode may oxidize. □

Received 8 February; accepted 9 June 1999.

- Ren, X., Wilson, M. S. & Gottesfeld, S. High performance direct methanol polymer electrolyte fuel cells. *J. Electrochem. Soc.* **143**, L12–L14 (1996).
- Wang, J., Wasmus, S. & Savinell, R. F. Evaluation of ethanol, 1-propanol, and 2-propanol in a direct oxidation polymer-electrolyte fuel cell. A real-time mass spectrometry study. *J. Electrochem. Soc.* **142**, 4218–4224 (1995).
- Sfeir, J., Van herle, J. & McEvoy, A. J. in *Proc. 3rd European Solid Oxide Fuel Cell Forum* (ed. Stevens, P.) 267–276 (European Fuel Cell Forum, Switzerland, 1998).
- Putna, E. S., Stubenrauch, J., Vohs, J. M. & Gorte, R. J. Ceria-based anodes for the direct oxidation of methane in solid oxide fuel cells. *Langmuir* **11**, 4832–4837 (1995).
- Aida, T., Abudala, A., Ihara, M., Komiyama, H. & Yamada, K. in *Proc. 4th Int. Symp. on Solid Oxide Fuel Cells* (eds Dokiya, M., Yamamoto, O., Tagawa, H. & Singhal, S. C.) 801–809 (Electrochemical Soc., Pennington, 1995).
- Fung, K.-Z., Chen, J., Tanner, C. & Virkar, A. V. in *Proc. 4th Int. Symp. on Solid Oxide Fuel Cells* (eds Dokiya, M., Yamamoto, O., Tagawa, H. & Singhal, S. C.) 1018–1027 (Electrochemical Soc., Pennington, 1995).
- Tsai, T. & Barnett, S. A. Increased solid-oxide fuel cell power density using interfacial ceria layers. *Solid State Ionics* **98**, 191–196 (1997).
- Janssen, G. J. M., DeJong, J. P. & Huijsmans, J. P. P. in *Proc. 2nd European Solid Oxide Fuel Cell Forum* (ed. Thorstensen, B.) 163–172 (European SOFC Forum, Switzerland, 1996).
- Hendriksen, P. V. in *Proc. 5th Int. Symp. on Solid Oxide Fuel Cells* (eds Stimming, U., Singhal, S. C., Tagawa, H. & Lehnert, W.) 1314–1325 (Electrochemical Soc., Pennington, 1997).
- Ray, E. R. in *1992 Fuel Cell Seminar Abstr.* (eds Fuel Cell Seminar Committee) 415–418 (Courtesy Associates, Washington, 1992).
- Huebner, W., Anderson, H. U., Reed, D. M., Sehl, S. R. & Deng, X. in *Proc. 4th Int. Symp. on Solid Oxide Fuel Cells* (eds Dokiya, M., Yamamoto, O., Tagawa, H. & Singhal, S. C.) 696–705 (Electrochemical Soc., Pennington, 1995).
- deSouza, S., Visco, S. J. & DeJonghe, L. C. Thin-film solid oxide fuel cell with high performance at low-temperature. *Solid State Ionics* **98**, 57–61 (1997).
- Minh, N. Q. in *Proc. 4th Int. Symp. on Solid Oxide Fuel Cells* (eds Dokiya, M., Yamamoto, O., Tagawa, H. & Singhal, S. C.) 138–145 (Electrochemical Soc., Pennington, 1995).
- Godickemeier, M., Sasaki, K. & Gauckler, L. J. in *Proc. 4th Int. Symp. on Solid Oxide Fuel Cells* (eds Dokiya, M., Yamamoto, O., Tagawa, H. & Singhal, S. C.) 1072–1081 (Electrochemical Soc., Pennington, 1995).
- Tsai, T., Perry, E. & Barnett, S. Low-temperature solid-oxide fuel cells utilizing thin bilayer electrolytes. *J. Electrochem. Soc.* **144**, L130–L132 (1997).
- Tsai, T. & Barnett, S. A. Effect of mixed-conducting interfacial layers on solid oxide fuel cell anode performance. *J. Electrochem. Soc.* **145**, 1696–1701 (1998).
- Steele, B. C. H., Middleton, P. H. & Rudkin, R. A. Material science aspects of SOFC technology with special reference to anode development. *Solid State Ionics* **40/41**, 388–393 (1990).
- Hori, C. E. et al. Thermal stability of oxygen storage properties in a mixed CeO₂–ZrO₂ system. *Appl. Catal. B* **16**, 105–117 (1998).

Acknowledgements. E.P.M. was supported by the Illinois Minority Graduate Incentive Program.

Correspondence and requests for materials should be addressed to S.A.B. (e-mail: s-barnett@nwu.edu).

Chronology, causes and progression of the Messinian salinity crisis

W. Krijgsman*, F. J. Hilgen†, I. Raffi‡, F. J. Sierro§ & D. S. Wilson||

* Paleomagnetic Laboratory "Fort Hoofddijk", Utrecht University, Budapestlaan 17, 3584 CD Utrecht, The Netherlands

† Department of Geology, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, The Netherlands

‡ Dipartimento di Scienze della Terra, Università "G. D'Annunzio", Campus Universitario, Via dei Vestini 31, 66013 Chieti Scalo, Italy

§ Department de Geologia, Facultad de Ciencias, Universidad de Salamanca, 37008 93106, Salamanca, Spain

|| Department of Geological Sciences, University of California, Santa Barbara, California 93106, USA

The Messinian salinity crisis is widely regarded as one of the most dramatic episodes of oceanic change of the past 20 or so million years (refs 1–3). Earliest explanations were that extremely thick evaporites were deposited in a deep and desiccated Mediterranean basin that had been repeatedly isolated from the Atlantic Ocean^{1,2}, but elucidation of the causes of the isolation—whether driven largely by glacio-eustatic or tectonic processes—have been hampered by the absence of an accurate time frame. Here we present an astronomically calibrated chronology for the Mediterranean Messinian age based on an integrated high-resolution stratigraphy and 'tuning' of sedimentary cycle patterns to variations in the Earth's orbital parameters. We show that the onset of the Messinian salinity crisis is synchronous over the entire Mediterranean basin, dated at 5.96 ± 0.02 million years ago. Isolation from the Atlantic Ocean was established between 5.59 and 5.33 million years ago, causing a large fall in Mediterranean water level followed by erosion (5.59–5.50 million years ago) and deposition (5.50–5.33 million years ago) of non-marine sediments in a large 'Lago Mare' (Lake Sea) basin. Cyclic evaporite deposition is almost entirely related to circum-Mediterranean climate changes driven by changes in the Earth's precession, and not to obliquity-induced glacio-eustatic sea-level changes. We argue in favour of a dominantly tectonic origin for the Messinian salinity crisis, although its exact timing may well have been controlled by the ~400-kyr component of the Earth's eccentricity cycle.

Most hypotheses about the initiation of the Messinian salinity crisis (MSC) agree that it resulted from a complex combination of tectonic and glacio-eustatic processes which progressively restricted and finally isolated the Mediterranean Sea from the open ocean^{1–8}. The gradual modification of water exchange with the Atlantic caused important palaeoceanographic changes in the Mediterranean. This is reflected in the classic Messinian sequence of Sicily⁹ which starts at 7.24 Myr ago (ref. 3) with alternations of open marine marls and sapropels, passes via diatomites into the Lower Evaporites (gypsum, evaporitic limestone and halite), and ends, above an erosional surface, with the Upper Evaporites (gypsum, marls) and fresh to brackish water deposits of Lago Mare facies. Here we define the MSC as the interval of evaporite deposition and Lago Mare sedimentation in the Mediterranean before the Pliocene flooding 5.33 Myr ago¹⁰.

Controversies still exist, however, over the timing and duration of the MSC; these range from a synchronous event¹ (that is, onset of the MSC in all basins at the same time) to a two-step event⁴ (onset of MSC first in marginal basins and later in deep basins) to a completely diachronous evolution⁵ (onset of MSC totally dependent on local basinal setting). Equally large controversies exist over

the cause, and the effects, of the isolation of the Mediterranean; the two basic explanations are (1) a large glacio-eustatic sea-level drop, related to expanding polar ice volume⁶, and (2) orogenic uplift accompanied by gravity-driven sliding of large nappe complexes in the Gibraltar arc⁷. Until now, correlations of stable-isotope ($\delta^{18}\text{O}$ and $\delta^{13}\text{C}$) records from open-ocean sequences to the Messinian event stratigraphy of the Mediterranean have been ambiguous because of the absence of a reliable time frame for the MSC. The establishment of astronomical polarity timescales for the past 10 Myr (refs 3, 11) provided a significant advance in dating the geological record and promised a solution for the MSC controversies. Unfortunately, the Mediterranean-based astronomical polarity timescale showed a gap during much of the Messinian (6.7–5.3 Myr ago)³, related to the presence of less-favourable sediments and the notoriously complex geological history of the Mediterranean in this time interval. However, the classic Messinian sediments display distinct sedimentary cyclicity, holding great promise for astronomical dating. Cyclostratigraphic and detailed palaeoclimatic studies revealed that the sedimentary cycles of the pre-evaporites are dominantly controlled by precession-induced changes in circum-Mediterranean climate^{3,12,13}.

To obtain a high-resolution cyclostratigraphic framework for the Messinian, we subjected three continuous pre-evaporite sequences of the western (Sorbas basin of Spain), central (Caltanissetta basin of Sicily) and eastern (Gavdos basin of Greece) Mediterranean to a detailed integrated stratigraphic study. All our stratigraphic records start at the Tortonian/Messinian boundary, and extend into the evaporites or evaporitic limestones of the MSC. Cyclostratigraphic correlations between the sections are straightforward, and have been confirmed by high-resolution biostratigraphy (Fig. 1). New magnetostratigraphic results were obtained from the Sorbas basin.

The starting point of our astronomical calibration is the previously established 'tuning' of early Messinian (up to 6.7 Myr ago) sequences³. Upward tuning, by calibrating younger cycles to successive insolation peaks¹⁴, generally shows a good to excellent fit between the characteristic sedimentary cycle patterns and the astronomical target curve (Fig. 1), in particular the precession/obliquity interference patterns in the insolation curve. Alternating thick/thin beds consistently correlate with high/low amplitude variations in insolation, proving that no sedimentary cycles are missing and that alternative correlations can be ruled out. Nevertheless, some minor misfits can also be observed, which are probably due to small inaccuracies in the applied astronomical solution. Continuing research directed at improving the accuracy of the astronomical solution may result in minor modifications in the future, but this would not seriously affect the tuning and, hence, the Messinian astrochronology.

The two normal polarity intervals in the Sorbas magnetostratigraphy undoubtedly correspond to chrons C3An.1n and C3An.2n, respectively. Most recent timescales have dealt with poor age-control in the Messinian by recalibrating the marine magnetic anomaly ages for C3An (from Cande and Kent¹⁵) to match an age of ~5.23 Myr for the oldest Pliocene reversal^{13,16}. Our astronomically tuned ages for the palaeomagnetic reversals of C3An, which are significantly older than in previous timescales (Table 1), close an important gap in the astronomical calibration.

Our magnetostratigraphy is confirmed by the position of the sinistral/dextral coiling change of *Neoglobobadrina acostaensis* in the middle of chron C3An.1r, in agreement with results from DSDP Site 609¹⁷ of the adjacent Atlantic. Confirmation of our ages comes from open-ocean calcareous nannofossil biochronology. ODP Site 853 in the equatorial Pacific¹⁸ has a reliable nannofossil biostratigraphy and magnetostratigraphy, but lacks a reliable astronomical tuning for the Miocene. ODP Site 926 in the equatorial Atlantic has reliable nannofossil events¹⁹ and a straightforward astronomical tuning²⁰, but lacks a magnetostratigraphy. Assuming a low-latitude global synchronicity of the nannofossil events and exporting their

astronomical ages to the Pacific site allows us to calculate the ages of the palaeomagnetic reversals of C3An. These ages are in agreement with our ages (Table 1), indicating that both astronomical time-scales are consistent with one another and confirming our age model for the Mediterranean Messinian. The age of the C3An.1n(y) reversal is especially well constrained because the 6.02-Myr LCO *Amaurolithus amplificus* datum occurs at 38.55–38.45 m in Hole 835B, indistinguishable from the 38.45-m depth of the reversal (Table 1). (Here LCO indicates last common occurrence.)

Support for our older ages for chron C3An is available from their implications for rates of sea-floor spreading. Detailed measurements of distances of tectonic-plate motion based on marine magnetic anomalies have provided a powerful test of the calibration

of the Plio-Pleistocene reversal chronology²¹, as spreading rates are often constant for intervals of 4 Myr or longer. Significant tectonic changes complicate the interpretation of the latest Miocene, but our revised ages provide the simplest picture of these changes (Fig. 2). Our timescale implies a constant rate of motion for Australia relative to Antarctica since 7 Myr ago, with a change then from 64 mm yr⁻¹ to 69 mm yr⁻¹ for a flowline at 98° E. Changes for Pacific Ocean plates are more dramatic, and because rates on different plate pairs change by different ratios, no possible timescale can yield constant rates for all plate pairs. Our ages yield the satisfying result that the largest rate change for each Pacific plate pair occurred simultaneously for all pairs at approximately 5.9 Myr. Previous age estimates¹⁶ implied more complicated tectonic histories, generally

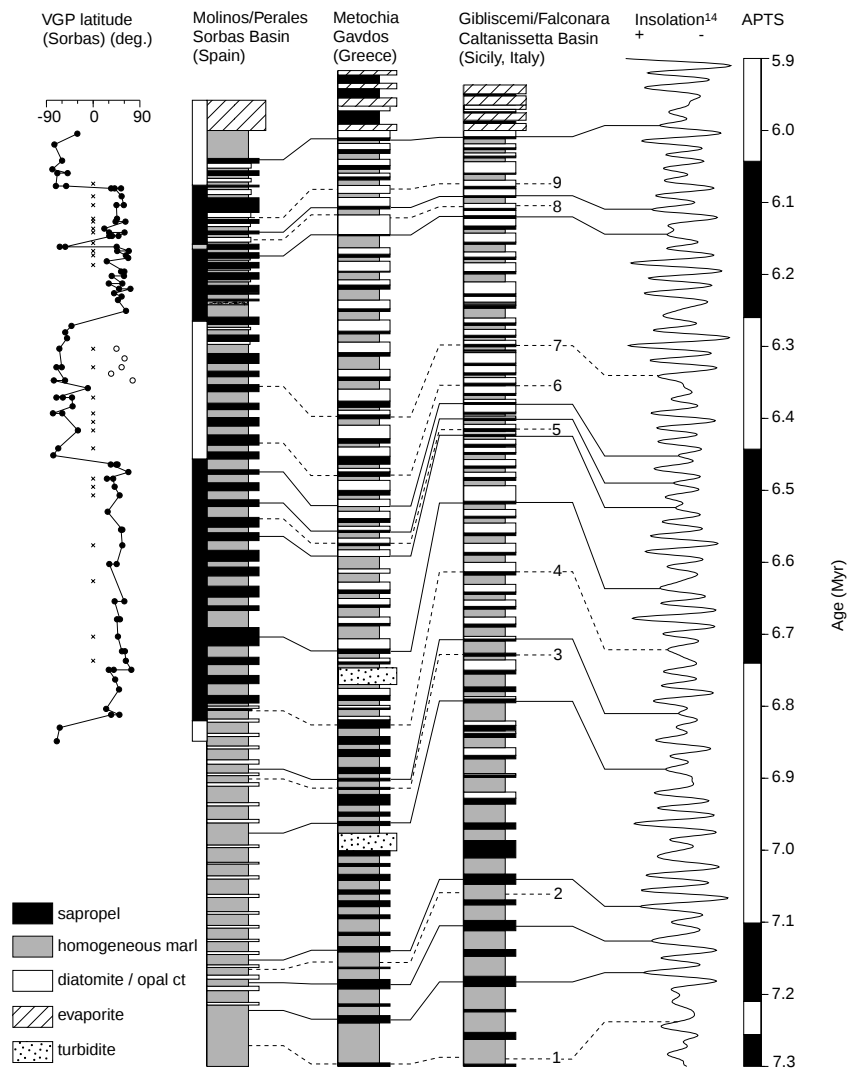


Figure 1 Astronomical calibration of Messinian pre-evaporite sequences. We used the La90 solution¹⁴ with present-day (1,1) values for the dynamical ellipticity of the Earth and the tidal dissipation by the Moon as target curve, because it is the most accurate solution from a geological point of view¹⁰. Total thickness of the pre-evaporite section in Sorbas is 125 m, on Gavdos 31 m, and on Sicily 50 m. Sedimentary cyclicity in the Sorbas basin is characterized by an alternation of homogeneous marls and opal ct-rich beds in the lower part and of sapropels and homogeneous marls with irregularly intercalated diatomites in the upper part. Cyclicity on Gavdos and Sicily is bipartite (sapropel/marls) in the lower part and mainly tripartite (sapropel/diatomite/marl) in the upper part. Laminites (sapropels and diatomites) correspond to insolation maxima, homogeneous marls to insolation minima, with the exception of the lower part of Sorbas¹³. Magnetostratigraphic data from Sorbas are represented as virtual geomagnetic pole (VGP) latitude.

Filled symbols denote samples with a palaeomagnetic signal sufficiently strong (0.01–0.1 mA m⁻¹) to determine the polarity. Open symbols denote directions interpreted as secondary overprint, crosses denote unreliable results. Cyclostratigraphic correlations are confirmed in detail by high-resolution planktonic foraminiferal biostratigraphy. Nine Mediterranean-wide planktonic foraminiferal bioevents are recognised: 1, first regular occurrence (FRO) of *Globorotalia conomiozea* group; 2, last occurrence (LO) of dominantly sinistral *Globorotalia scitula*; 3, *Globorotalia nicolae* FO; 4, *G. nicolae* LO; 5, *G. conomiozea* group LO; 6, first common occurrence (FCO) of *Turborotalita multiloba*; 7, sinistral/dextral coiling change *Neogloboquadrina acostaensis*; 8, first influx sinistral neogloboquadrinids (90%); 8, second influx sinistral neogloboquadrinids (40%). APTS (astronomical polarity timescale) is based on Sorbas data supplemented by earlier results from Crete³.

involving faster apparent rates (more positive slopes) near 5.5 Myr ago and slower apparent rates near 7–8 Myr ago, as if the age of C3A is artificially young. Simultaneous rate changes by varying amounts on different Pacific plate pairs could easily be a consequence of a significant change in the absolute motion of the Pacific plate, as has been suggested by numerous authors for this time interval; for example, Cox and Engebretson²². Two studies have independently determined ages of 5.8–5.9 Myr (relative to C3An.1n(y) at 5.94–5.95 Myr) for a clockwise shift in motion direction of the Pacific plate relative to the Juan de Fuca plate²³ and the Antarctic plate²⁴. According to the ages we determine here, the times of rate changes and direction changes are indistinguishable. As earlier shown for the Pliocene²¹, it is evident that the astronomical polarity timescale for the late Miocene results in a more realistic spreading-rate history

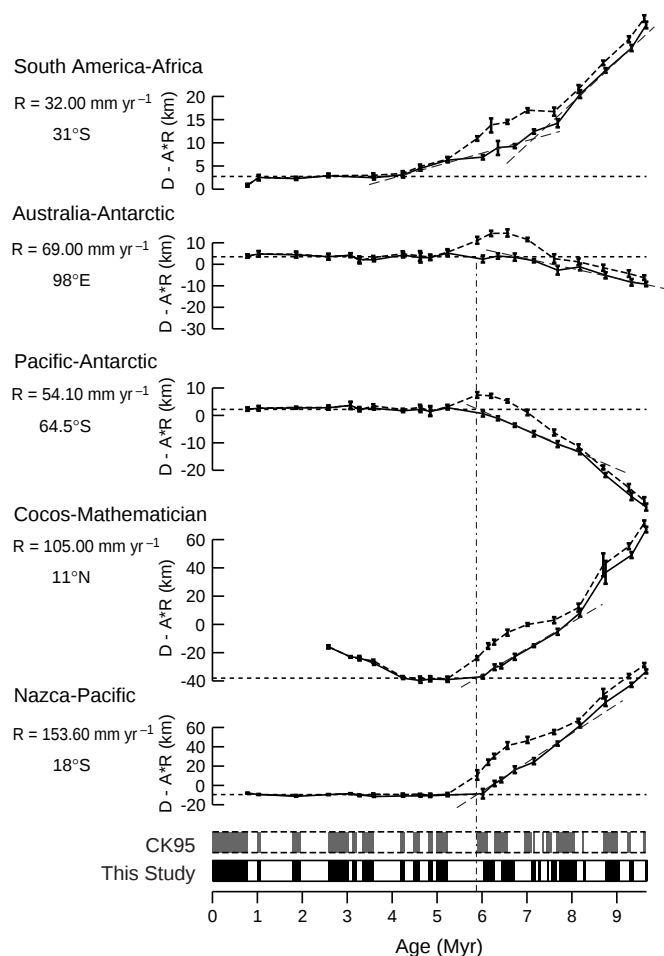


Figure 2 Late Neogene sea-floor spreading-rate histories. Reduced spreading distance versus age is shown for five plate pairs with intermediate-to-fast spreading rate and good data coverage. Reduced distance is simply the observed distance minus the predicted distance for a constant spreading rate. Distance scales are plotted inversely to the spreading rates so that, for plate pairs spreading constantly at the reduction rate, timescale errors will plot as uniform vertical departures from the reduction line. Our revised ages (solid lines) imply constant Australia-Antarctic and South America-Africa motion for 4–7 Myr and simultaneous changes on Pacific ocean plate pairs at 5.9 Myr (vertical line). All distance measurements are based on new determinations of rotation poles using digital data from the NGDC archive and miscellaneous recent cruises. Error bars are 95% confidence intervals for distance. Cocos-Mathematician data are from north of the Clipperton fracture zone, $10\text{--}13^\circ\text{N}$ only; Pacific-Antarctic data are restricted to new data from $63\text{--}66^\circ\text{S}$ ²⁴; Australia-Antarctic data from $94\text{--}138^\circ\text{E}$; Nazca-Pacific data from $1^\circ\text{N}\text{--}30^\circ\text{S}$, and South America-Africa data from $25\text{--}29^\circ\text{S}$ compiled by C. Weiland³² span the well mapped and demonstrably rigid areas of the plates. D, distance; A, age; R, rate.

than does the most widely used geomagnetic polarity timescale (GPTS) of Cande and Kent¹⁶.

The Messinian astrochronology that we report here solves many of the existing controversies over the MSC. The cyclostratigraphic results show that the transition to evaporite deposition (Lower Evaporites) occurs at the same sedimentary cycle in all sections. This proves that the MSC is a perfectly synchronous event over the entire Mediterranean, the onset of which is dated astronomically at 5.96 ± 0.02 Myr ago while the total duration is approximately 640 kyr. Deposition of the Lower Evaporite unit is thus independent of the palaeogeographic and geodynamic setting of the individual basins. The Lower Evaporites of Sorbas and Caltanissetta both require a continuously marine environment, excluding a relative sea-level fall exceeding the palaeodepth of these basins (that is, <200 m; ref. 25). This favours a deep-water model instead of a shallow-water (repetitive process of desiccating and reflooding) model for the deep ($>1,000$ m) Mediterranean basins. Complete isolation and possible desiccation was only established after deposition of the Lower Evaporites, when the Mediterranean water level dropped more than 1,000 m as shown by incised canyons of the Rhone, Ebro, Po and Nile rivers in the Mediterranean margins⁴. Deposition of the Upper Evaporite unit, overlying erosional surfaces, took place in a non-marine, deep Mediterranean basin forming a large Lago Mare²⁶. Canyon incision in the Aegean region may have caused the transition to Lago Mare conditions by capturing the Black Sea drainage^{1,26}.

Our astronomically calibrated timescale for the Messinian now allows us to relate global environmental signals to the Mediterranean events. The most prominent (± 50 m) glacio-eustatic sea-level falls (at oxygen isotope stages TG22 and TG20²⁷), which are often suggested to have triggered the MSC^{4,6}, clearly post-date the onset of evaporite deposition by 200–260 kyr and pre-date the isolation event by 100–160 kyr. Therefore, obliquity-controlled, glacio-eustatic lowering of sea level may only have added a minor contribution to the restriction and isolation processes. On the contrary, the onset of the MSC clearly coincides with the amplitude increase in insolation following the ~ 400 -kyr eccentricity minimum dated around 6.0–6.1 Myr, suggesting that long-term orbital cycle forcing superimposed on a directional trend (for example, tectonic closure) plays a critical role in the exact timing of the event. The same may also hold for the onset of the Upper Evaporites (~ 5.6 Myr ago), as well as for other intra-Messinian events such as the Tortonian/Messinian (T/M) boundary event (~ 7.2 Myr ago) and the marked change in lithology in the Sorbas basin around 6.7–6.8 Myr ago.

The timing of the changes in plate-tectonic spreading rates (Fig. 2) indicate that tectonic processes indeed contributed to the onset of the MSC, but a direct relation is not obvious. Additional evidence for tectonic activity before the MSC comes from the two pre-Messinian Atlantic gateways. Orogenic uplift had already closed the Betic corridor through Spain in the latest Tortonian/earliest Messinian^{7,28}, and severely constricted the Rifian corridor through Morocco in the earliest Messinian²⁹. The exact timing of closure of the entire Rifian corridor is less certain, because large olistolith complexes cover the youngest marine sediments in the central parts. Mammal exchange between Spain and Morocco—signifying a near-landbridge—had already occurred at least 6.1 Myr ago, 140 kyr before the onset of evaporite deposition²⁸. These tectonic processes obviously limited the water exchange between the Mediterranean and the Atlantic during the Messinian, and thus probably heralded the onset of the MSC.

Field evidence, showing that marl/sapropel alternations pass upward via diatomites into evaporite/sapropel alternations, indicates that the sedimentary cyclicity in the Lower (16–17 cycles; refs 8, 25) and Upper (7–8 cycles; ref. 30) Evaporites is controlled by astronomical precession. The total number of sedimentary cycles is in good agreement with the total number of precession peaks,

Table 1 Age control for magnetic reversals

Reversal	Astronomical age (Sorbas basin)	HKLLSZ ⁹	CK95 ¹⁶	SCHPS ¹¹	CK92 ¹⁵	Constant spreading rate age
C3An.1n (y)	6.04 ± 0.01	5.952	5.894	5.875	5.705	6.01
C3An.1n (o)	6.26 ± 0.02	6.214	6.137	6.122	5.946	6.28
C3An.2n (y)	6.44 ± 0.01	6.356	6.269	6.256	6.078	6.43
C3An.2n (o)	6.71 ± 0.03†	6.677	6.567	6.555	6.376	6.73
Nannofossil event	Age (Site 926) ¹⁸	Depth (m.c.d.) (Hole 853B)	Reversal	Depth (m.c.d.) (Hole 853B)	Interpolated age (nannofossils)	
1. LO <i>Discoaster quinqueramus</i>	5.537	32.67	C3An.1n (y)	38.45	6.02	
2. LCO <i>Amaurolithus amplifucus</i>	6.020	38.50	C3An.1n (o)	NA	NA	
3. XA. <i>amplifucus</i> *, <i>T. rugosus</i>	6.798	47.80	C3An.2n (y)	43.60	6.45	
4. FOA. <i>amplifucus</i> *	6.840	48.25	C3An.2n (o)	46.50	6.69	

Comparison of our astronomical polarity reversal ages with the ages according to recently advanced polarity timescales of Hilgen *et al.*², Shackleton *et al.*¹¹ and Cande and Kent^{15,16} (upper Table). Error estimates in our astronomical ages refer to the uncertainty in the exact stratigraphic position of the magnetic polarity reversal with respect to the astronomically dated cycles. † Denotes average of the astronomical ages from Sorbas and Crete². The new astronomical ages are in good to excellent agreement with independently derived ages that result from the most constant spreading rates (see also Fig. 2). Note that only a slight discrepancy exist in the ages of C3An.1n(y). Four correlatable nannofossil bioevents with astronomical age and depth from ODP Site 926 (Atlantic) and 853B (Pacific) are used to interpolate the age of the magnetic reversals from Site 853B (lower table; note that C3An.1n (o) was not recovered). Reversal ages are calculated by exporting astronomical ages of the nannofossil events from ODP Site 926. Ages are in Myr; m.c.d. indicates metres composite depth; X denotes cross-over between *Amaurolithus amplifucus* (*including transitional forms) and *Triquetrorhabdulus rugosus*; NA, not available.

whereas there is clearly not enough time for a 40-kyr obliquity control, thus excluding glacio-eustasy. Evaporite deposition occurred during precession maxima (insolation minima), during relatively dry periods when evaporation exceeded precipitation. This allowed the Mediterranean to form a giant brine, when outflow of dense Mediterranean waters into the Atlantic was restricted or obstructed by shallow sills in the western Mediterranean gateways. During precession minima (insolation maxima) and relatively wet periods, high freshwater runoff resulted in deposition of sapropel-like sediments. As the average periodicity for precession in Neogene times is 21.7 kyr, the Lower Evaporite and the Upper Evaporite units have a duration of approximately 370 kyr and 175 kyr, respectively. Tentatively calibrating the evaporite and post-evaporite cycles to the insolation curve leaves only a small 'Messinian gap' (between 5.59 and 5.50 Myr ago), during which the desiccation of the Mediterranean and the accompanying isostatic rebound processes³¹ (tectonic tilting and erosion) must have occurred. □

Received 5 November 1998; accepted 10 June 1999.

- Hsu, K. J., Ryan, W. B. F. & Cita, M. B. Late Miocene desiccation of the Mediterranean. *Nature* **242**, 240–244 (1973).
- Ryan, W. B. F. *et al.* *Initial Reports of the Deep Sea Drilling Project* Vol. 13 (US Govt Printing Office, Washington, 1973).
- Hilgen, F. J. *et al.* Extending the astronomical (polarity) time scale into the Miocene. *Earth Planet. Sci. Lett.* **136**, 495–510 (1995).
- Clauzon, G., Suc, J.-P., Gautier, F., Berger, A. & Loutre, M.-F. Alternate interpretation of the Messinian salinity crisis: Controversy resolved? *Geology* **24**, 363–366 (1996).
- Butler, R. W. H., Lickorish, W. H., Grasso, M., Pedley, H. M. & Ramberti, L. Tectonics and sequence stratigraphy in Messinian basins, Sicily: constraints on the initiation and termination of the Mediterranean 'salinity crisis'. *Geol. Soc. Am. Bull.* **107**, 425–439 (1995).
- Hodell, D. A., Benson, R. H., Kent, D. V., Boersma, A. & Rakic-El Bied, K. Magnetostratigraphic, biostratigraphic, and stable isotope stratigraphy of an Upper Miocene drill core from the Salé Briqueterie (northwest Morocco): A high-resolution chronology for the Messinian stage. *Paleoceanography* **9**, 835–855 (1994).
- Wijermars, R. Neogene tectonics in the Western Mediterranean may have caused the Messinian Salinity Crisis and an associated glacial event. *Tectonophysics* **148**, 211–219 (1988).
- Vai, G. B. in *Miocene Stratigraphy: An Integrated Approach* (eds Montanari, A., Odin, G. S. & Coccioni, R.) 463–476 (Elsevier Science, Amsterdam, 1997).
- Decima, A. & Wezel, F. Late Miocene evaporites of the Central Sicilian Basin. *Init. Rep. DSDP Leg 13*, 1234–1240 (1973).
- Lourens, L. J. *et al.* Evaluation of the Pliocene to early Pleistocene astronomical time scale. *Paleoceanography* **11**, 391–413 (1996).
- Shackleton, N. J., Crowhurst, S., Hagelberg, T., Pisias, N. G. & Schneider, D. A. A new late Neogene time scale: Application to leg 138 sites. *Proc. ODP Sci. Res.* **138**, 73–101 (1995).
- Sproveri, R., Di Stefano, E., Caruso, A. & Bonomo, S. High resolution stratigraphy in the Messinian Tripoli Formation in Sicily. *Paleopelagos* **6**, 415–435 (1996).
- Sierro, F. J. *et al.* Messinian pre-evaporite sapropels and precession-induced oscillations in western Mediterranean climate. *Mar. Geol.* **153**, 137–146 (1999).
- Laskar, J., Joutel, F. & Boudin, F. Orbital, precessional, and insolation quantities for the Earth from –20 Myr to +10 Myr. *Astron. Astrophys.* **270**, 522–533 (1993).
- Cande, S. C. & Kent, D. V. A new geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **97**, 13917–13951 (1992).
- Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**, 6093–6095 (1995).
- Hooper, P. W. P. & Weaver, P. P. E. Paleogeographic significance of Late Miocene to early Pliocene planktonic foraminifers at deep sea drilling project site 609. *Init. Rep. DSDP* **94**, 923–934 (1987).
- Raffi, I., Rio, D., d'Atri, A., Fornaciari, E. & Rochetti, S. Quantitative distribution patterns and biomagnetostратigraphy of middle and late Miocene calcareous nannofossils from equatorial Indian and Pacific Oceans (Legs 115, 130 and 138). *Proc. ODP Sci. Res.* **138**, 479–503 (1995).

- Backman, J. & Raffi, I. Calibration of Miocene nannofossil events to orbitally tuned cyclostratigraphies from Ceara Rise. *Proc. ODP Sci. Res.* **154**, 83–99 (1997).
- Shackleton, N. J. & Crowhurst, S. Sediment fluxes based on an orbitally tuned time scale 5 Ma to 14 Ma, Site 926. *Proc. ODP Sci. Res.* **154**, 69–82 (1997).
- Wilson, D. S. Confirmation of the astronomical calibration of the magnetic polarity timescale from sea-floor spreading rates. *Nature* **364**, 788–790 (1993).
- Cox, A. & Engebretson, D. C. Change in motion of the Pacific plate at 5 Ma. *Nature* **313**, 472–474 (1985).
- Wilson, D. S. Confidence intervals for motion and deformation of the Juan de Fuca. *J. Geophys. Res.* **98**, 16053–16071 (1993).
- Cande, S. C., Raymond, C. A., Stock, J. & Haxby, W. F. Geophysics of the Pittman Fracture Zone and Pacific-Antarctic plate motions during the Cenozoic. *Science* **270**, 947–953 (1995).
- Dronkert, H. Evaporite models and sedimentology of Messinian and Recent evaporites. *GUA Papers of Geology Series 1*, Vol. 24 (1985).
- McCulloch, M. T. & De Deckker, P. Sr isotope constraints on the Mediterranean environment at the end of the Messinian salinity crisis. *Nature* **342**, 62–65 (1989).
- Shackleton, N. J., Hall, M. A. & Plate, D. Pliocene stable isotope stratigraphy of Site 846. *Proc. ODP Sci. Res.* **138**, 337–355 (1995).
- Garcés, M., Krijgsman, W. & Agustí, J. Chronology of the late Turolian deposits of the Fortuna basin (SE Spain): implications for the Messinian evolution of the eastern Betics. *Earth Planet. Sci. Lett.* **163**, 69–81 (1998).
- Krijgsman, W. *et al.* Late Neogene evolution of the Taza-Guercif Basin (Rifian Corridor; Morocco) and implications for the Messinian salinity crisis. *Mar. Geol.* **153**, 147–160 (1999).
- Masce, G. & Heimann, K. O. Geological observations from Messinian and lower Pliocene outcrops in Sicily. *Mem. Soc. Geol. It.* **16**, 127–140 (1976).
- Norman, S. E. & Chase, C. G. Uplift of the shores of the western Mediterranean due to Messinian desiccation and flexural isostasy. *Nature* **322**, 450–451 (1983).
- Grindlay, N. R., Weiland, C. M. & Fox, P. J. *Eos* **76**, F573 (1995).

Acknowledgements. We thank C. G. Langereis, W.-J. Zachariasse and M.-F. Loutre for comments on the manuscript. This work was supported by GOA/NWO, NSF, DGICYT and Fundacion Arceles.

Correspondence and requests for materials should be addressed to W.K. (e-mail: krijgsma@geo.uu.nl).

The origin of snake feeding

Michael S. Y. Lee*, Gorden L. Bell Jr† & Michael W. Caldwell‡

* Department of Zoology, University of Queensland, St Lucia, Brisbane, Queensland 4072, Australia

† Museum of Geology, South Dakota School of Mines and Technology, Rapid City, South Dakota 57701, USA

‡ Paleobiology, Canadian Museum of Nature, PO Box 3443, Station 'D', Ottawa, Ontario K1P 6P4, Canada

Snakes are renowned for their ability to engulf extremely large prey, and their highly flexible skulls and extremely wide gape are among the most striking adaptations found in vertebrates^{1–5}. However, the evolutionary transition from the relatively inflexible lizard skull to the highly mobile snake skull remains poorly understood, as they appear to be fundamentally different and no obvious intermediate stages have been identified^{4,5}. Here we present evidence that mosasaurs—large, extinct marine lizards related to snakes—represent a crucial intermediate stage. Mosasaurs, uniquely among lizards, possessed long, snake-like palatal teeth for holding prey. Also, although they retained the rigid

upper jaws typical of lizards, they possessed highly flexible lower jaws that were not only morphologically similar to those of snakes, but also functionally similar. The highly flexible lower jaw is thus inferred to have evolved before the highly flexible upper jaw—in the macrophagous common ancestor of mosasaurs and snakes—for accommodating large prey. The mobile upper jaw evolved later—in snakes—for dragging prey into the oesophagus. Snakes also have more rigid braincases than lizards, and the partially fused meso- and metakinetic joints of mosasaurs are transitional between the loose joints of lizards and the rigid joints of snakes. Thus, intermediate morphologies in snake skull evolution should perhaps be sought not in small burrowing lizards, as commonly assumed, but in large marine forms.

Recent phylogenetic analyses^{6,7} indicate that the nearest relatives of snakes are the Mosasauroidae (here termed 'mosasaurs'), medium-sized to gigantic Cretaceous marine lizards. Mosasaurs include the medium-sized dolichosaurids and aigialosaurids as well as the large to gigantic mosasaurids^{8–10}. The most basal ('primitive') snakes are the Cretaceous forms *Pachyrhachis*^{6,11} and *Dinilysia*¹², along with living blindsnakes (scoleophidians) and anilioids (basal alethinophidians)^{4,13–15}. Identification of the nearest relatives of snakes and the most primitive snakes (Fig. 1) reveals transitional stages in the evolution of snake feeding.

Lizards possess mobile joints between the braincase and skull roof (the metakinetic joint) and across the posterior skull roof (the mesokinetic joint)^{16,17}. In snakes, both joints are fused, resulting in a rigid skull 'core' from which the highly flexible jaws are suspended^{1,5}. Mosasaurs, consistent with their phylogenetic posi-

tion between lizards and snakes (Fig. 1), had an intermediate morphology in which both joints were partially immobilized, with extensive tight overlaps, but not closed sutures^{18,19}.

Lizards have rather inflexible upper and lower jaws. In the upper jaw, the maxilla and palatal arch (palatine + pterygoid) are sutured to neighbouring elements, allowing, at most, only slight palatal movements⁵. The two lower jaw rami cannot move appreciably with respect to one another, being suturally united anteriorly at a firm symphysis. Within each lower jaw ramus, the anterior and posterior portions are also tightly integrated in a complex, overlapping contact, although varanoid lizards have a slightly looser arrangement²⁰.

Snakes have highly flexible upper and lower jaws. In the upper jaw, each maxilla and palatal arch can move independently. Some mobility is present in all snakes^{11,15,21–24} except leptotyphlopoid blindsnakes. These mobile upper jaws facilitate transport of (often large) prey by unilateral jaw movements, associated with 'snout shifting' in basal snakes²² and elaborated in advanced snakes as the 'pterygoid walk', where alternate movements of the left and right palatal arches drag the skull over the prey, forcing it into the oesophagus^{23,24}.

The lower jaws are also highly mobile. In the basal snake *Cylindrophis*²², the symphyseal tips are rounded, permitting hinge-like movements, but they cannot separate. In addition, a loose intramandibular joint divides each lower jaw ramus into anterior and posterior halves. Most of this joint consists of flexible connective tissue; however, ventrally there is a tight vertical bony contact between the angular and splenial. This contact is deep but

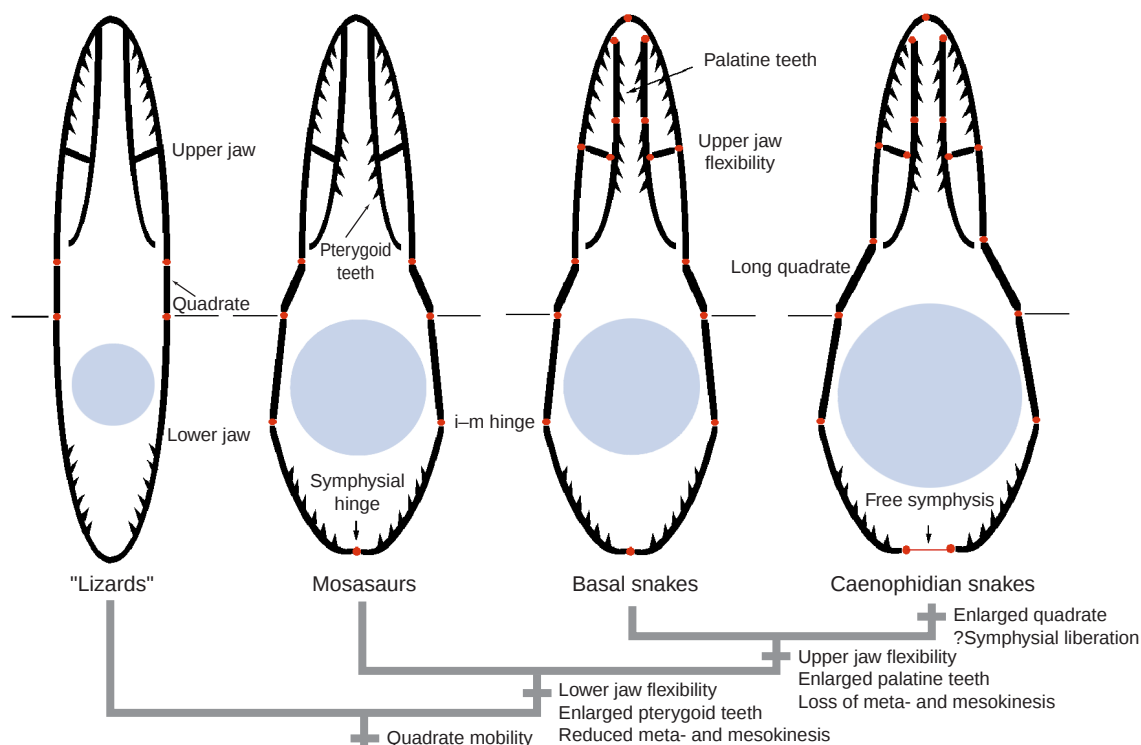


Figure 1 The evolution of the highly flexible snake jaw system, superimposed on a simplified cladogram of snake relationships^{4–6,14}. 'Basal snakes' include *Pachyrhachis*, *Dinilysia*, scoleophidians, anilioids, *Xenopeltis* and *Loxocemus*. Booid-grade snakes, not shown, appear to be intermediate between basal snakes and caenophidians in possessing symphyseal liberation but lacking an enlarged quadrate. In each diagram, the upper jaw, quadrate and lower jaws are portrayed as if opened 180° (as in ref. 1). Red regions denote mobile joints, black bars denote bony units. The evolutionary changes leading from the rigid lizard morphology to the flexible snake morphology are indicated on the cladogram nodes. From left to right across the skull diagrams, note (1) increase in gape size,

(2) gradual increase in the number of mobile joints, (3) lower-jaw joints evolving before upper-jaw joints, so that (4) mosasaurs possess very snake-like lower jaws but retain lizard-like upper jaws. The distribution of extensive symphyseal liberation is equivocal. Histology indicates it is absent in most basal snakes (for example, scoleophidians, *Xenopeltis*) and present in most macrostomatan snakes (such as booids, pythons, acrochordids and most colubrids)²⁵. However, apparent exceptions exist, and functional studies are required (for example, histology indicates that basal snake *Cylindrophis* has extensive mandibular liberation²⁵, but this is contradicted by observations on living animals²²).

narrow, and there are two vertical ridges from the angular that fit into corresponding grooves on the splenial (Fig. 2f). The overall morphology permits extensive mediolateral (horizontal) flexion, but restricts dorsoventral (vertical) movements. The symphyseal and intramandibular hinges function together to increase the distance between the lower jaw rami ('gape'), allowing ingestion of larger or wider prey (Fig. 1). Horizontal flexion at the intramandibular joint causes each jaw ramus to bulge laterally. This in turn causes the anterior tips of the two rami to form a more obtuse angle with respect to each other, such movements being accommodated by the flexible symphyseal hinge²². The ventral ends of the quadrates, too, can splay out slightly to increase the diameter of the posterior mouth²⁴, an ability found in all squamates¹⁷. The lower jaws in all other primitive snakes (except some blindsnakes; see later) are very similar morphologically to *Cylindrophis* and, from observations of living animals, are also similar functionally^{11,21,22}. Examination of osteological and cleared-and-stained specimens (listed in ref. 6) indicates that the intramandibular hinge is also present in leptotyphlopids, the basal blindsnake lineage^{4,15}, although the symphysis might not be very mobile²⁵. In contrast, other blindsnakes (typhlopids and anomalepidids) have consolidated the intramandibular hinge,

but retain the symphyseal hinge^{4,25,26}. These changes might be correlated with increasing specialization on numerous small prey (ants and termites)¹⁵, and (in typhlopids and anomalepidids) heavy reliance on the highly mobile maxilla for feeding^{15,26}. The presence of both functionally related hinges in *Pachyrhachis*, *Dinilysia*, anilioids and some blindsnakes indicates that they are primitive for snakes. More advanced snakes retain both joints, but many have elaborated this system to increase gape further. In some, the mandibular tips can separate appreciably^{1-3,22,24,25}; although functional studies across the relevant taxa have not been done, morphological studies indicate that extensive 'mandibular liberation' has a complex distribution within alethinophidians, and might not simply be a uniquely caenophidian innovation²⁵. Mandibular liberation, in addition to increasing gape, facilitates unilateral movements of the lower jaw²⁴. In caenophidians, the quadrates are elongated, resulting in more effective quadrate 'flaring'²⁴ (Fig. 1).

Mosasauroids, as the nearest relatives of snakes, are here suggested to be morphologically intermediate between other squamates (lizards) and snakes in terms of jaw kinesis. However, rather than having incipient snake-like kinesis in both jaws, as might be expected, they retained primitive lizard-like upper jaws while possessing snake-like lower jaws (Fig. 1). These lower jaws were present in all three groups of mosasaurs: dolichosaurids, aigialosaurids and mosasaurids. Because all relatively complete specimens of dolichosaurs and aigialosaurs are preserved on slabs and cannot be manipulated, here we discuss the mosasaur condition based on a large number of disarticulated, three-dimensionally preserved mosasaurid jaw elements²⁷ that reveal the mobility and function at various joints.

In mosasaurs the upper jaws (maxilla and palatal elements) remained relatively rigid^{8,19}. The maxilla was firmly sutured to the premaxilla, prefrontal, jugal, palatine and ectopterygoid. The palatine was firmly sutured to the vomer. The pterygoid was tightly sutured to the ectopterygoid, but the pterygoid-palatine and ectopterygoid-jugal contacts were looser. As in lizards, and unlike snakes, the maxilla, palatine and (probably) pterygoid were capable of only limited movement. However, uniquely among lizards, mosasaurs possessed long, recurved palatal teeth (Fig. 1). Many basal snakes have similar teeth (*Pachyrhachis*, *Anilius*, *Cylindrophis*, basal uropeltids and some *Dinilysia*), and phylogenetic analyses (refs 6, 11 and J. Scanlon and M. Lee, unpublished data) indicate that such teeth may be primitively present in snakes, with losses in scolecophidians and within the uropeltid-*Anomochilus* clade. Bite marks indicate that in mosasaurs, as in snakes, prey was bitten initially with the maxillary teeth and later held with these palatal teeth²⁷. In snakes, the system is further elaborated so that the palatal teeth are used not just to hold prey, but to 'walk' the skull over it, thus dragging it into the oesophagus^{23,24}.

In contrast to the upper jaw, which remained very lizard-like, the mosasaur lower jaw was extremely flexible and snake-like (Figs 1, 2). As in snakes, the symphysis was loose and mobile: each ramus terminated anteriorly in a smoothly rounded tip^{8,10}. Bite marks indicate that the mandibular tips could change their angle with respect to each other²⁷. These marks also indicate that the mandibular tips (and thus the two jaw rami) never separated during feeding. Thus, as in primitive snakes, the mobile symphysis in mosasaurs appears to have functioned as a hinge.

Mosasaurs also possessed an extremely mobile and snake-like intramandibular joint in the middle of each jaw ramus (Fig. 2a-e). It was much more flexible than the irregular, overlapping contact characteristic of other squamates, including varanid lizards. Although most previous interpretations of the mosasaur intramandibular joint suggested that it permitted mainly dorsoventral movements^{8,18,27,28}, the morphology of the joint surfaces indicates that it was designed instead for mediolateral flexion. Such motion has been suggested previously²⁹, but without compelling supporting evidence.

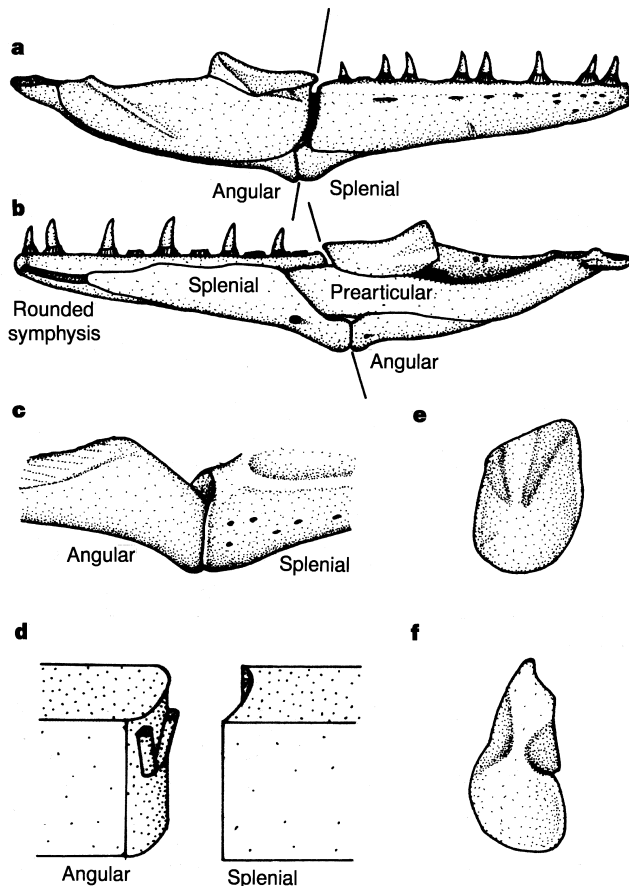


Figure 2 The lower jaw of a mosasaur, *Platecarpus*. **a**, Lateral and **b**, medial views, showing the intramandibular joint (vertical line). Note the loose contacts between the anterior and posterior jaw portions, except in the ventral region. **c**, Close-up of the ventral (splenial-angular) contact, in lateral view. **d**, Enlarged diagram of the intramandibular joint in mosasaurs. The joint in primitive snakes is very similar, except that the curvature of the joint in the transverse axis is less. Posterior view of the right splenial, which forms the anterior articular surface of the intramandibular joint, in the mosasaur *Platecarpus* (**e**) and the basal snake *Cylindrophis* (**f**). Lateral is to the right. Note the similarities, such as the deep but narrow joint surface and two vertical grooves. Scale: **a**, **b**, $\times 0.14$; **c**, $\times 0.4$; **e**, $\times 0.6$; **f**, $\times 30$ (**a**, **c**, **e**, after ref. 8; **b**, after ref. 10; **f**, after BMNH 1947.1.1.8).

For convenience, the mosasaur intramandibular joint will be described as if the lower jaw was orientated as a vertical sheet of bone, as is usually drawn (Fig. 2a–d). In life, this sheet was slightly inclined ventromedially. Hence, the mediolateral hinging movements described below—orthogonal to the plane of the lower jaw—will have a slight dorsoventral component.

The morphology of the mosasaur intramandibular joint was very similar to that in basal snakes such as *Cylindrophis* (Fig. 2f). The anterior and posterior halves of the lower jaw were separated by a large fissure filled with connective tissue, and there was a close bony contact only in the ventral region (Fig. 2a, b). This contact was deep but narrow. The posterior articulatory surface, formed by the angular, was smoothly convex in the transverse dimension but almost flat in the vertical dimension, and bore two converging ridges (Fig. 2c–e). These contours matched the anterior articulatory surface, formed by the splenial, which was concave transversely but flat vertically, and bore two converging grooves. The groove-and-ridge system did not permit translational (sliding) movements in any direction. As the vertical axes of the joint surfaces are flat and extensive, vertical hinging (folding) movements would have resulted in wide separation (dislocation) of large areas of the joint. However, extensive horizontal hinging movements were permitted because the horizontal axes of the joint surfaces were short and smoothly rounded. This mediolateral hinging movement causes the anterior and posterior portions of each lower jaw to change their angle with respect to each other in the horizontal plane. The rest of the vertical contact between the anterior and posterior jaw portions was loose and filled with a large region of connective tissue (Fig. 2a, b), permitting extensive horizontal movement. Also, the anterior flange of the prearticular (which was loosely sandwiched by the dentary–splenial unit) was a thin, vertical sheet (Fig. 2b) which could bend appreciably in the horizontal but not the vertical plane. Manipulation of undistorted, three-dimensionally prepared bones of several mosasaur genera reveals that the anterior portion can swing through a substantial arc around the posterior portion, from 22° to 37° (see Methods). This is more than the movement possible in skeletal material of extant lizards, where the range is at most 15° (in *Varanus*)²⁰, and usually much less. Without information from living animals and soft tissue, the movement range possible in live mosasaurs cannot be ascertained. However, because the bones in mosasaurs allow much more movement than the bones in all other lizards, and because the joint morphology closely resembles the highly mobile junction in basal snakes, it can be reasonably concluded that the intramandibular joint was much more flexible in mosasaurs than in other lizards.

The two joints in the mosasaur lower jaw thus seem to have been similar both morphologically and functionally to the corresponding two joints in primitive snakes such as *Cylindrophis* (Fig. 2e, f). They interacted geometrically in the same way, increasing gape by bulging outwards at the intramandibular hinge, with such movements being accommodated by the symphyseal hinge (Fig. 1). However, although generally very similar and functionally equivalent, mosasaur lower jaws differed from those of snakes in some respects, such as absence of the anterolateral process of the compound (surangular) projecting into the dentary^{8–10}. As in snakes²⁴, the quadrates in mosasaurs were capable of splaying outwards to increase the diameter of the posterior mouth^{18,19}. These features for increasing gape are consistent with the observation that mosasaurs had narrow snouts, common in aquatic tetrapods that need to reduce water resistance to rapid mouth closure. At the same time, they were the top predators in Cretaceous seas and consumed extremely large prey such as sea turtles and other marine reptiles³⁰. Macrophagy also characterizes the primitive snake *Pachyrhachis* (found with remains of a large fish in its gut¹¹), and nearly all other snakes (the highly modified scolecophidians being a notable exception). *Pachyrhachis* also has a relatively slender snout. Thus, the common ancestor of mosasaurs and snakes was macrophagous and (less certainly)

slender-snouted, and evolved increased mobility at the intramandibular joint as an adaptation for accommodating large prey. However, highly flexible upper jaws did not evolve until much later, and are restricted to snakes. Mosasaurs thus represent a transitional morphology, possessing flexible lower jaws as advanced as those of basal snakes, but simultaneously retaining the immobile upper jaws of lizards.

Although the rigid braincase and highly flexible jaws of modern snakes represent a highly integrated functional system, the presence of some but not all of these features in mosasaurs indicates that they did not evolve simultaneously (Fig. 1). The reduction of metakinesis and mesokinesis and development of large palatal dentition evolved early and also characterize mosasaurs. The highly flexible lower jaw also evolved early and is present in mosasaurs; however, the highly flexible upper jaw evolved later and is restricted to snakes alone. Most recent studies (reviewed in refs 4–6, 11, 15) have speculated that small, limbless burrowing lizards are the most plausible analogues for snake ancestors. However, the above observations support a very different model. In terms of skull structure, the large, limbed marine mosasaurs were functionally, as well as phylogenetically, intermediate between lizards and snakes. □

Methods

The anterior and posterior jaws were placed in the maximally 'straightened' position, when further lateral movement of the anterior jaw was impeded by contacts with the posterior jaw elements, especially the prearticular and angular. A line was drawn along the medial edge, from the intramandibular joint to the preserved anterior tip of each jaw or splenial fragment. The posterior part of the jaw was then held rigid and its anterior was moved into the maximally 'flexed' position, when further medial movement of the anterior jaw was impeded by the bony contacts with the posterior jaw (again, principally the prearticular and angular). A second line was then drawn along the medial edge of the flexed anterior element. The internal angle between the two lines was measured with a protractor. The same technique was used with extant lizard material, so the results should be comparable. Note that this method might underestimate the movement possible in mosasaur joints compared with extant lizards, as fossilized bones cannot bend whereas those of extant lizards are slightly flexible.

The disarticulated mosasauroid specimens used in the manipulations, and the observed range of movement for each, are as follows [Abbreviations: YPM, Peabody Museum, Yale University; NMC, Canadian Museum of Nature; SDSM, Museum of Geology, South Dakota School of Mines.]: *Clidastes liodontus* (YPM 24914), 37°; *Mosasaurus maximus* (YPM 430), 22°; *Mosasaurus conodon* (SDSM 26172), 24°; *Platecarpus cf. somenensis* (YPM 40734), 37°; *Platecarpus* sp. (NMC 40914), 33°; *Plioplatecarpus primaevus* (SDSM 40596), 35°; *Plioplatecarpus primaevus* (SDSM 40597), 36°.

Received 6 January; accepted 18 May 1999.

- Gans, C. The feeding mechanism of snakes and its possible evolution. *Am. Zool.* **1**, 207–227 (1961).
- Greene, H. W. Dietary correlates of the origin and radiation of snakes. *Am. Zool.* **23**, 431–441 (1983).
- Cundall, D. In *Snakes: Ecology and Evolutionary Biology* (eds Seigel, R. A., Collins, J. T. & Novak, S. S.) 106–140 (Macmillan, New York, 1987).
- Rieppel, O. A review of the origin of snakes. *Evol. Biol.* **22**, 37–130 (1988).
- Kardong, K. V., Kiene, T. L. & Bels, V. Evolution of trophic systems in squamates. *Neth. J. Zool.* **47**, 411–427 (1997).
- Lee, M. S. Y. Convergent evolution and character correlation in burrowing reptiles: towards a resolution of squamate phylogeny. *Biol. J. Linn. Soc.* **65**, 369–453 (1998).
- Caldwell, M. W. Squamate phylogeny and the relationships of snakes and mosasauroids. *Zool. J. Linn. Soc.* **125**, 115–147 (1999).
- Russell, D. A. Systematics and morphology of American mosasaurs. *Bull. Peabody Mus. Nat. Hist.* **23**, 1–237 (1967).
- Bell, G. L. Jr in *Ancient Marine Reptiles* (eds Nicholls, E. L. & Callaway, J.) 293–332 (Academic, New York, 1997).
- DeBraga, M. & Carroll, R. L. The origin of mosasaurs as a model of macroevolutionary patterns and processes. *Evol. Biol.* **27**, 245–322 (1993).
- Lee, M. S. Y. & Caldwell, M. W. Anatomy and relationships of *Pachyrhachis problematicus*, a primitive snake with hindlimbs. *Phil. Trans. R. Soc. Lond. B* **353**, 1521–1552 (1998).
- Estes, R., Frazzetta, T. H. & Williams, E. E. Studies on the fossil snake *Dinilysia patagonica* Woodward: Pt 1. Cranial morphology. *Bull. Mus. Comp. Zool. Harv.* **140**, 25–74 (1970).
- Underwood, G. A contribution to the classification of snakes. *Br. Mus. (Nat. Hist.) Publ.* **653**, 1–179 (1967).
- Cundall, D., Wallach, V. & Rossman, D. A. The systematic relationships of the snake genus *Anomochilus*. *Zool. J. Linn. Soc.* **109**, 275–299 (1993).
- Greer, A. *The Biology and Evolution of Australian Snakes* (Surrey Beatty, Sydney, 1997).

16. Frazetta, T. H. The origin of amphikinesis in lizards: a problem in functional morphology and the evolution of adaptive systems. *Evol. Biol.* **20**, 419–461 (1986).
17. Arnold, E. N. Cranial kinesis in lizards: variations, uses, and origins. *Evol. Biol.* **30**, 323–357 (1998).
18. Callison, G. Intracranial mobility in Kansas mosasaurs. *Paleontol. Contr. Univ. Kansas* **26**, 1–15 (1967).
19. Lingham-Soliar, T. Anatomy and functional morphology of the largest marine reptile known, *Mosasaurus hoffmani* (Mosasauridae, Reptilia) from the Upper Cretaceous, Upper Maastrichtian of The Netherlands. *Phil. Trans. R. Soc. Lond. B* **347**, 155–180 (1995).
20. Pregill, G. K., Gauthier, J. A. & Greene, H. W. The evolution of helodermatid squamates, with description of a new taxon and an overview of Varanoidea. *Trans. San Diego Nat. Hist. Soc.* **21**, 167–202 (1986).
21. Frazetta, T. H. Studies on the fossil snake *Dinilysia patagonica* Woodward. II. Jaw machinery in the earliest snakes. *Forma Functio* **3**, 205–221 (1970).
22. Cundall, D. Feeding behaviour in *Cylindrophis* and its bearing on the evolution of alethinophidian snakes. *J. Zool.* **237**, 353–376 (1995).
23. Bolt, R. E. & Ewer, R. F. The functional anatomy of the head of the puff adder, *Bitis arietans* (Merr.). *J. Morphol.* **114**, 83–106 (1964).
24. Kardong, K. V. Kinematics of swallowing in the yellow rat snake, *Elaphe obsoleta quadrivittata*: a reappraisal. *Jap. J. Herpetol.* **11**, 96–109 (1986).
25. Young, B. A. The comparative morphology of the intermandibular connective tissue in snakes (Reptilia: Squamata). *Zool. Anz.* **237**, 59–84 (1998).
26. Iordansky, N. N. Jaw apparatus and feeding mechanics of *Typhlops* (Ophidia: Typhlopidae): a reconsideration. *Russ. J. Herpetol.* **4**, 120–127 (1997).
27. Kauffman, E. G. & Kesling, R. V. An upper Cretaceous ammonite bitten by a mosasaur. *Contr. Mus. Paleont. Univ. Michigan* **15**, 193–248 (1960).
28. Russell, D. A. Intracranial mobility in mosasaurs. *Postilla* **86**, 1–19 (1964).
29. Gregory, J. T. Convergent evolution: the jaws of *Hesperornis* and the mosasaurs. *Evolution* **5**, 345–354 (1951).
30. Massare, J. A. Tooth morphology and prey preference of Mesozoic marine reptiles. *J. Vert. Paleontol.* **7**, 121–137 (1987).

Acknowledgements. We thank V. Wallach, G. Underwood, K. Karding, N. Kley, J. Scanlon and A. Greer for discussion and comments on the manuscript; J. Rosado and J. Cadle (Museum of Comparative Zoology, Harvard University), O. Rieppel and A. Resat (Field Museum of Natural History), C. Holton, T. Trombone and L. Ford (American Museum of Natural History), S. Chapman and C. McCarthy (British Museum of Natural History) and B. Purdy and K. de Queiroz (National Museum of Natural History, Smithsonian Institution) for access to materials under their care; and the Australian Research Council, Fulbright Foundation and NERC (Canada) for funding.

Correspondence should be addressed to M.L. (e-mail: mlce@zoology.uq.edu.au).

Planktonic nutrient regeneration and cycling efficiency in temperate lakes

Jeff J. Hudson, William D. Taylor* & David W. Schindler

Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada

* Department of Biology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

Planktonic nutrient regeneration is a fundamental process that maintains most of the primary productivity in marine and freshwater environments. However, there is no robust predictive model to describe the pattern and efficiency of nutrient cycling across aquatic systems. Based on rather weak evidence, the efficiency of nutrient regeneration is believed to decline along a gradient of productivity, so that nutrient-poor environments are assumed to be more efficient at cycling their nutrients than are nutrient-rich environments^{1–5}. Here we measure phosphorus regeneration directly and show that cycling efficiency does not vary with phosphorus concentration. In addition, we confirm that the phosphorus supply for lake plankton comes primarily from within the plankton community, rather than from external loading or from larger organisms such as fish.

To evaluate nutrient cycling efficiency, we measured phosphorus regeneration rates from 20 temperate freshwater lakes from three major physiographical regions of North America (Table 1). These lakes provided a nutrient gradient of 4–80 $\mu\text{g l}^{-1}$ of total phosphorus (P), spanning almost the entire range found at northern latitudes, and a trophic gradient that would include lakes of a wide range of productivities⁶. Regeneration rates of dissolved P from the entire planktonic community were measured by using a radioisotope technique⁷.

Phosphorus regeneration increased linearly with total P (Fig. 1). When the variates (total P and regeneration rate) were transformed to provide equal variance across the range in total P (Fig. 1, inset), 96% of the variation in regeneration rate was accounted for by total P ($\log_{10}$ regeneration rate ($\text{ng P l}^{-1} \text{ h}^{-1}$) = $1.0077(\log_{10}$ total P

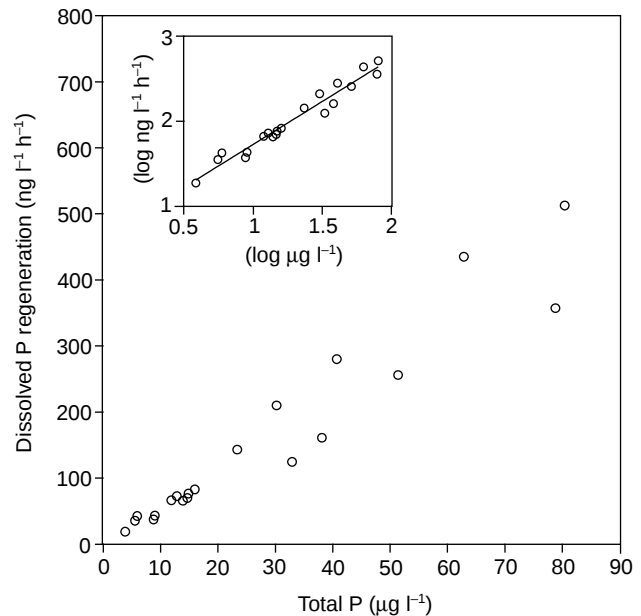


Figure 1 Planktonic regeneration of dissolved P ($\text{ng P l}^{-1} \text{ h}^{-1}$) as a function of total P ($\mu\text{g l}^{-1}$). The inset displays the same relationship with transformed values ($\log_{10}$) of regeneration and total P ($n = 20$, $R^2 = 0.96$, $\log_{10}$ regeneration rate ($\text{ng P l}^{-1} \text{ h}^{-1}$) = $1.0077(\log_{10}$ total P ($\mu\text{g l}^{-1}$)) + 0.7206).

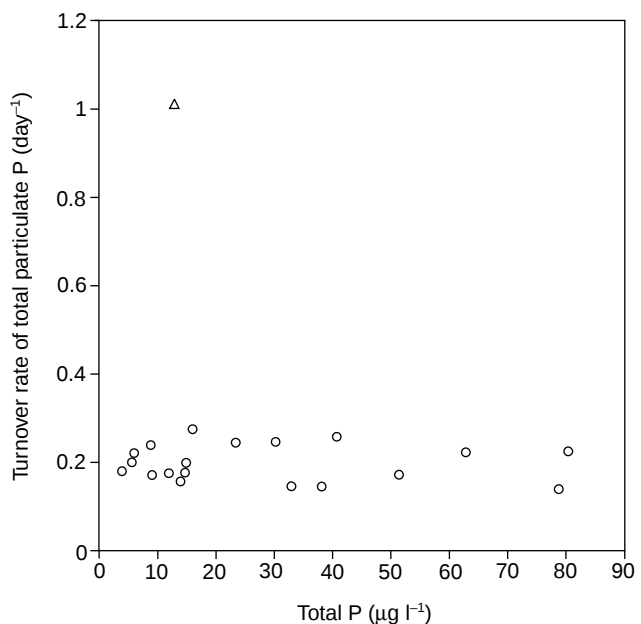


Figure 2 Cycling efficiency or turnover rate (per day) of particulate phosphorus (circles) as a function of total phosphorus ($\mu\text{g l}^{-1}$). Particulate phosphorus (P that collects on a filter of pore size $<0.2 \mu\text{m}$) is an approximate measure of the biomass of plankton in a sample of water. Cycling efficiency or turnover rate is equal to the regeneration rate divided by particulate phosphorus. Mean turnover rate was 20% per day with a range of 14 to 28%. Lake Hibernia (triangle) was considered to be an outlier and was not included in the analyses. Unlike other lakes, Lake Hibernia water (20 l) was carried for 3 km down a mountainside. This handling may have disrupted the plankton community—for example, only 13% of the total P was in the particulate P fraction. In other lakes, the average particulate P concentration was 66% of total P, with a range of 43 to 81%.

($\mu\text{g P l}^{-1}$) + 0.7206). This relation indicates that regeneration by plankton is of primary importance: for example, Mirror Lake in New Hampshire, whose nutrient budget has been extensively studied⁸, has a mean summer total P concentration of $\sim 5 \mu\text{g P l}^{-1}$ in its surface waters, which would equate to a regeneration rate of $639 \text{ ng P l}^{-1} \text{ d}^{-1}$ with our regeneration equation. The mean daily P input (from the atmosphere, hypolimnetic mixing, surface runoff and other forms) into the epilimnion of Mirror Lake has been estimated at $115 \text{ ng P l}^{-1} \text{ d}^{-1}$ during July and August⁹ and only represents 18% of the daily estimate for planktonic regeneration.

For decades, aquatic scientists have debated the importance of nutrient regeneration by fish (compare, for example, refs 10–14 with 15–17). Now we can compare phosphorus regeneration by fish with phosphorus regeneration by plankton. Two studies that have concluded that P regeneration by fish is important will serve for this comparison. Red-breasted sunfish (*Lepomis auritus*) were added to a mesocosm experiment¹¹ in Wolf Lake at a biomass equivalent to all species in the lake. Regeneration by these fish was $\sim 40 \text{ ng P l}^{-1} \text{ d}^{-1}$, or less than 4% of planktonic P regeneration ($1,025 \text{ ng P l}^{-1} \text{ d}^{-1}$, using the equation in Fig. 1 and a total P value of $8 \mu\text{g l}^{-1}$). Gizzard shad (*Dorosoma cepedianum*) were the dominant fish in eutrophic Acton Lake¹³. Gizzard shad P excretion was estimated to be $1,400 \text{ ng P l}^{-1} \text{ d}^{-1}$, or less than 11% of planktonic regeneration ($13,000 \text{ ng P l}^{-1} \text{ d}^{-1}$, using the equation in Fig. 1 and a total P value of $99 \mu\text{g l}^{-1}$). Therefore, fish appear to make a relatively small contribution to total P regeneration compared with estimated planktonic regeneration.

The linear relationship between total P and P regeneration

indicates that the paradigm that regeneration is more efficient in oligotrophic (low productivity) systems is incorrect. Instead, our results indicate that P regeneration per unit of particulate phosphorus (PP; that is, the rate constant for P regeneration) does not vary with total phosphorus. If regeneration is expressed as a fraction of the PP pool (that is, as a turnover rate), then regeneration is about 20% per day across the entire trophic gradient (range, 14–28%; Fig. 2). In other words, regeneration releases an amount of phosphorus equal to the particulate phosphorus pool every 5 days (range, 3.5–7 days).

This outcome is consistent with other measures of nutrient turnover. It has been found¹⁸ that herbivores graze the same proportion of annual net primary production in nutrient-poor and in nutrient-rich aquatic systems. In another comparative study, the proportion of planktonic primary production lost to sinking of particles decreased with increasing primary production in 15 North American lakes¹⁹ (chlorophyll *a* range, $0.45\text{--}18.6 \mu\text{g l}^{-1}$); if oligotrophic systems are more efficient in cycling their nutrients, then the reverse trend (that is, a greater proportion of particle sedimentation in eutrophic than in oligotrophic systems) should have been observed. The authors concluded that the cycling of nutrients within the epilimnia (as opposed to export by sedimentation) was not more efficient in oligotrophic lakes¹⁹.

The nutrient content of lakes, measured as total P, determines many aspects of lake ecosystem function^{6,20,21}. Our results show that it is nutrient regeneration, and not cycling efficiency or turnover rate of the particulate pool, that varies with total P. In these lakes, there is about twofold variation in turnover rate around a mean of

Table 1 Lake characteristics

Lake	Landform and location	Latitude, longitude	Date sampled*	Area (Ha)	Maximum depth (m)	Mixing depth (m)†	Total P ($\mu\text{g P l}^{-1}$)	Additional information (ref.)
Halfmoon	Interior Plains, Alberta	53° 24' N, 113° 04' W	150797	41	8.5	4.5	80	23
Nakamun	Interior Plains, Alberta	53° 53' N, 114° 12' W	210797	354	8.0	‡	79	23
Astotin	Interior Plains, Alberta	53° 41' N, 112° 51' W	180897	562	7.0	‡	63	24
Teen§	Interior Plains, Alberta	54° 29' N, 113° 43' W	090997	36	8.4	5.5	23	25
Jenkins	Interior Plains, Alberta	54° 45' N, 113° 35' W	090997	90	13	5	30	25
Bilsky§	Interior Plains, Alberta	54° 45' N, 113° 35' W	090997	8.3	9.0	3	33	25
Schumaker§	Interior Plains, Alberta	54° 45' N, 113° 35' W	090997	20	6.8	2	41	25
Ghost	Interior Plains, Alberta	54° 45' N, 113° 35' W	090997	95	13	6.5	51	25
Cameron	Interior Plains, Alberta	53° 38' N, 114° 01' W	130997	12	7	5	38	NA
Heart	Interior Plains, Saskatchewan	53° 59' N, 106° 02' W	030997	56	5.2		14	26
Namekus	Interior Plains, Saskatchewan	53° 50' N, 106° 02' W	030997	773	27	8.5	15	26
Trapper	Interior Plains, Saskatchewan	53° 48' N, 106° 01' W	030997	227	4.6		15	26
Trefoil 2	Montane, Alberta	52° 54' N, 118° 03' W	230897	1.8	17	8.5	3.8	27
Patricia	Montane, Alberta	52° 54' N, 118° 06' W	230897	69	42	8.5	8.8	27
Trefoil 1	Montane, Alberta	52° 54' N, 118° 03' W	230897	3.5	10	4.5	9.0	27
Cabin	Montane, Alberta	52° 53' N, 118° 08' W	230897	32	20	3	12	27
Hibernia	Montane, Alberta	52° 52' N, 118° 08' W	230897	10	9.0	NA	13	27
Leach	Montane, Alberta	52° 42' N, 117° 54' W	230897	8.5	11	8	16	27
Ranger	Precambrian Shield, Ontario	45° 09' N, 78° 51' W	1993–1995	11	13	2.6¶	5.5¶	28
Mouse	Precambrian Shield, Ontario	45° 11' N, 78° 51' W	1993–1995	9.0	9.0	3.0¶	5.9¶	28

* As day, month and year.

† To the nearest half metre.

‡ Temperature declined gradually without a distinct mixing depth (epilimnion).

§ Lakes without official names; authors assigned the name.

|| Isothermal.

¶ Mean of three summers (1993–95).

NA, Not available or not measured.

20% per day. It remains to be determined whether this variation can be attributed to food-web structure, the ratio of nutrients (C:N:P), or to environmental factors such as light and temperature. □

Methods

Field sampling. Methods specific to Mouse and Ranger Lake are described elsewhere⁷. Methods for the remaining 18 lakes are described below. Only lakes that had a maximum depth ≥ 4 m were considered, in order to minimize benthic effects on the pelagic zone (Table 1). Lakes on rivers were avoided, to limit the effects of river water on the pelagic environment. Lakes with minimal shoreline development (except Nakamun Lake) were chosen, to minimize human effects on lake nutrient cycling. Most lakes had thermally stratified water columns; however, two isothermal lakes were included from the Interior Plains (Table 1). Only lakes with total phosphorus concentrations below $100 \mu\text{g l}^{-1}$ were selected. Water (20 litres) was removed at the mid-epilimnetic depth of each lake with a Van Dorn sampler and placed in 20-litre polyethylene containers (acid-washed) held in coolers. When a distinct epilimnion or mixing depth was not present, water was removed from just below the surface (≤ 1 m).

Laboratory methods. Detailed laboratory methods are described elsewhere⁷. Lake water was placed in 4-litre clear polyethylene containers that had been washed (0.1% contrad-70), rinsed (ethanol) and leached (0.1 M HCl). Each sample was incubated with carrier-free radiophosphate ($^{33}\text{PO}_4$; 270–2,100 Bq ml $^{-1}$; ICN Biomedicals) for ~ 27 h (range 21–47 h) to label the planktonic community. Lake water was incubated near ambient temperatures ($\pm 2^\circ\text{C}$), which ranged from 18 to 22°C . This range minimized the effects of temperature on rate measurements. Incubations were terminated by injection of competitive inhibitor ($^{31}\text{PO}_4$; final concentration 1–5 mg P litre $^{-1}$). This prevented re-incorporation of ^{33}P after it was released from the plankton. We then measured the accumulation of ^{33}P in the dissolved pool over time: the slope of this line provided an estimate of the release rate of dissolved ^{33}P . The remaining lake water was analysed for total P (ref. 22), which was calculated as the sum of dissolved and particulate P. The release rate of dissolved P was calculated by using the following formula: P release rate = ^{33}P release rate $\times$ total P/total ^{33}P , so our definition for phosphorus regeneration was the transfer of phosphorus from the particulate pool ($>0.2 \mu\text{m}$) to the dissolved pool ($<0.2 \mu\text{m}$) over time. Egestion, excretion, decay, cell lysis, cellular exudate and sloppy feeding (un-ingested food) all contribute to this process. Radioactivity was measured by liquid scintillation and corrected for background radioactivity. Quenching of samples was not detected.

Received 8 February; accepted 21 June 1999.

- Goldman, J. C. in *Flows of Energy and Materials in Marine Ecosystems* 1st edn, vol. 13 (ed. Fasham, M. J. R.) 137–170 (Plenum, New York, 1984).
- Sheldon, R. W. Phytoplankton growth rates in the tropical ocean. *Limnol. Oceanogr.* **29**, 1342–1346 (1984).
- Capblancq, J. Nutrient dynamics and pelagic food web interactions in oligotrophic and eutrophic environments: an overview. *Hydrobiologia* **207**, 1–14 (1990).
- Harris, G. P. *Phytoplankton Ecology: Structure, Function and Fluctuation* 1st edn (Chapman and Hall, London, 1986).
- Harris, G. P. Pattern, process and prediction in aquatic ecology. A limnological view of some general ecological problems. *Freshwat. Biol.* **32**, 143–160 (1994).
- Wetzel, R. G. *Limnology* 2nd edn (Saunders, New York, 1983).
- Hudson, J. J. & Taylor, W. D. Measuring regeneration of dissolved phosphorus in planktonic communities. *Limnol. Oceanogr.* **41**, 1560–1565 (1996).
- An Ecosystem Approach to Aquatic Ecology: Mirror Lake and its Environment* 1st edn (ed. Likens, G. E.) (Springer, New York, 1985).
- Caraco, N. F., Cole, J. J. & Likens, G. E. New and recycled primary production in an oligotrophic lake: insights for summer phosphorus dynamics. *Limnol. Oceanogr.* **37**, 590–602 (1992).
- Kraft, C. E. Phosphorus regeneration by Lake Michigan alewives in the mid-1970s. *Trans. Am. Fish. Soc.* **122**, 749–755 (1993).
- Perezfuentetaja, A., McQueen, D. J. & Ramcharan, C. W. Predator-induced bottom-up effects in oligotrophic systems. *Hydrobiologia* **317**, 163–176 (1996).
- Vanni, M. J. in *Food Webs: Integration of Patterns and Dynamics* 1st edn (eds Polis, G. A. & Winemiller, K. O.) 81–95 (Chapman and Hall, New York, 1996).
- Schaus, M. H. et al. Nitrogen and phosphorus excretion by detritivorous gizzard shad in a reservoir ecosystem. *Limnol. Oceanogr.* **42**, 1386–1397 (1997).
- Horpilla, J. et al. Top-down or bottom-up effects by fish: issues of concern in biomanipulation of lakes. *Restoration Ecol.* **6**, 20–28 (1998).
- Nakashima, B. S. & Leggett, W. C. The role of fish in the regulation of phosphorus availability in lakes. *Can. J. Fish. Aquat. Sci.* **37**, 1540–1549 (1980).
- den Oude, P. J. & Gulati, R. D. Phosphorus and nitrogen excretion rates of zooplankton from the eutrophic Loosdrecht lakes, with notes on other P sources for phytoplankton requirements. *Hydrobiologia* **169**, 379–390 (1988).
- Mazumder, A. et al. Partitioning and fluxes of phosphorus: mechanisms regulating the size-distribution and biomass of plankton. *Arch. Hydrobiol. Beih.* **35**, 121–143 (1992).
- Cyr, H. & Pace, M. L. Magnitude and patterns of herbivory in aquatic and terrestrial ecosystems. *Nature* **361**, 148–150 (1993).

- Baines, S. B. & Pace, M. L. Relationship between suspended particulate matter and sinking flux along a trophic gradient and implications for the fate of planktonic primary production. *Can. J. Fish. Aquat. Sci.* **51**, 25–36 (1994).
- Schindler, D. W. Evolution of phosphorus limitation in lakes. *Science* **195**, 260–262 (1977).
- Edmondson, W. T. *The Uses of Ecology: Lake Washington and Beyond* 1st edn (Univ. Washington Press, Seattle, 1991).
- Parsons, T. R., Maita, Y. & Lalli, C. M. *A Manual of Chemical and Biological Methods for Seawater Analysis* 1st edn (Pergamon, Oxford, 1984).
- Atlas of Alberta Lakes* 1st edn (eds Mitchell, P. & Prepas, E.) (Univ. of Alberta Press, Edmonton, 1990).
- Lin, C. K. *Phytoplankton Succession in Astotin Lake, Elk Island National Park, Alberta* (Thesis, Univ. Alberta, Edmonton, 1968).
- Gingras, B. A. & Paszkowski, C. A. Breeding patterns of common loons on lakes with three different fish assemblages in north-central Alberta. *Can. J. Zool.* (in the press).
- Prince Albert National Park Resource Description and Analysis* vol. 2 (Canadian Park Service) (Environment Canada, Parks, Prairie and Northern Region, Winnipeg, Manitoba, 1986).
- Anderson, S. R. Crustacean plankton communities of 340 lakes and ponds in and near the national parks of the Canadian Rocky Mountains. *J. Fish. Res. Board. Can.* **31**, 855–869 (1974).
- Ramcharan, C. W. et al. A comparative approach to determining the role of fish predation in structuring limnetic ecosystems. *Archiv. Hydrobiol.* **133**, 389–416 (1995).

Acknowledgements. We thank T. Paul, J. Almond, S. Leung, T. MacDonald, L. Lawton, B. Rolseth and B. Parker for field and laboratory assistance; D. McQueen and the Dorset Research Centre for logistical support at Mouse and Ranger Lakes; D. Watters, D. Donald, C. Paszkowski, B. Gingras, P. Mitchell and D. Zell for help with lake selection; and M. Pace, J. Murie and G. M. Taylor for helpful criticisms of the manuscript. This work was supported by a scholarship (NSERC, Canada) and a Killam postdoctoral fellowship (University of Alberta) to J.J.H., and NSERC operating grants to W.D.T. and D.W.S.

Correspondence and requests for materials should be addressed to J.J.H. (e-mail: jhudson@gpu.srv.ualberta.ca).

Genome complexity, robustness and genetic interactions in digital organisms

Richard E. Lenski*, Charles Ofria†, Travis C. Collier‡ & Christoph Adami§

* Center for Microbial Ecology, Michigan State University, East Lansing, Michigan 48824, USA

† Computation and Neural Systems and § Kellogg Radiation Laboratory, California Institute of Technology, Pasadena, California 91125, USA

‡ Department of Organismic Biology, Ecology, and Evolution, University of California, Los Angeles, California 90095, USA

Digital organisms are computer programs that self-replicate, mutate and adapt by natural selection^{1–3}. They offer an opportunity to test generalizations about living systems that may extend beyond the organic life that biologists usually study. Here we have generated two classes of digital organism: simple programs selected solely for rapid replication, and complex programs selected to perform mathematical operations that accelerate replication through a set of defined ‘metabolic’ rewards. To examine the differences in their genetic architecture, we introduced millions of single and multiple mutations into each organism and measured the effects on the organism’s fitness. The complex organisms are more robust than the simple ones with respect to the average effects of single mutations. Interactions among mutations are common and usually yield higher fitness than predicted from the component mutations assuming multiplicative effects; such interactions are especially important in the complex organisms. Frequent interactions among mutations have also been seen in bacteria, fungi and fruitflies^{4–6}. Our findings support the view that interactions are a general feature of genetic systems^{7–9}.

Many fundamental questions in biology are difficult to address, as a consequence of the high dimensionality of genomes^{7–9} as well as the practical difficulties of manipulating numerous genotypes and analysing their resulting phenotypic properties. Progress has been made using microorganisms^{4,5,10–18}, but these problems remain daunting. An alternative approach involves studying artificial life,

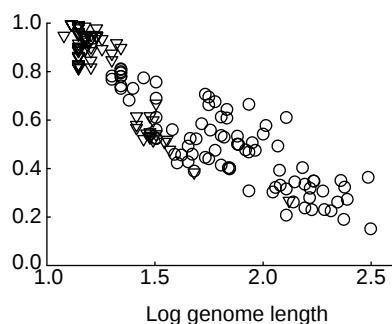


Figure 1 Proportion of single point mutations that are lethal for digital organisms. Shown as a function of $\log_{10}$ -transformed genome length. Circles, complex organisms; triangles, simple organisms.

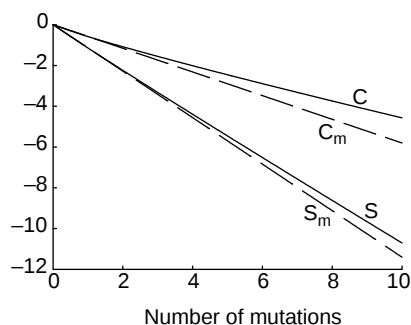


Figure 2 $\log_{10}$ -transformed mean fitness as a function of number of point mutations for simple and complex classes of digital organisms. Solid curve C shows the fitness function calculated using the average parameter values from the complex organisms. Dashed line C_m is their fitness function expected under the multiplicative model (obtained using the average α and setting $\beta = 1$). Solid curve S is the fitness function for the average simple organism. Dashed line S_m gives the corresponding function expected under the multiplicative model.

in particular certain computer programs—or digital organisms—that share with real organisms the properties of self-replication, mutation, competition and evolution, as well as genomes with high dimensionality and hence indeterminate evolutionary trajectories. The use of digital organisms to address biological questions is controversial, but it can be justified on several grounds. First, artificial life allows us to seek generalizations beyond the organic forms that biologists have studied to date, all of which derive from a common ancestor and share the same basic chemistry of DNA, RNA and proteins. Maynard Smith makes the case thus¹⁹: “So far, we have been able to study only one evolving system and we cannot wait for interstellar flight to provide us with a second. If we want to discover generalizations about evolving systems, we will have to look at artificial ones.” Second, digital organisms allow us to perform experiments on a scale that is unattainable with real organisms. Here we test billions of different genotypes to measure mutational effects and interactions; a recent experiment with the bacterium *Escherichia coli* did so using a few hundred genotypes⁴. Moreover, the performance of digital organisms can be measured exactly, whereas such data are subject to error, and hence loss of statistical power, in any real biological system. Third, there is growing interest in using programs that can evolve to solve complex computational problems^{20–23}. Knowing how mutations affect performance and interact with one another has important implications for setting parameters such as mutation and recombination rates, just as mutational effects and interactions can influence the evolution of these parameters in real organisms^{4,6,24–28}.

Our experiments were performed using Avida, a flexible platform for research on artificial life³. Briefly, digital organisms are self-

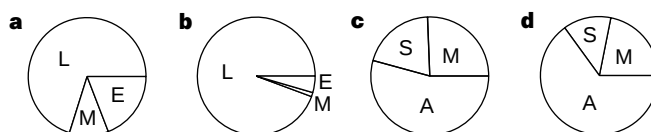


Figure 3 Proportions of mutational pairs classified according to their interaction. L, lethal: at least one mutation is lethal alone, as is the double mutant. M, multiplicative: neither mutation is lethal, and the relative fitness of the double mutant is exactly equal to the product of the relative fitnesses of the two single mutations. E, epistatic: the double mutant's fitness is unequal to the multiplicative expectation. S, synergistic: double mutant less fit than expected. A, antagonistic: the double mutant is more fit than expected. Average distributions are for complex organisms (a), simple organisms (b), complex excluding lethals (c), and simple excluding lethals (d).

Table 1 Comparisons between complex and simple digital organisms of genome size and several mutational-effect parameters

Response variable	Mean complex ($\pm$ s.d.)	Mean simple ($\pm$ s.d.)	Mean difference ($\pm$ s.d.)	P^*
Genome length	91.25 (69.07)	19.80 (14.18)	71.45 (64.76)	<0.0001
Decay test, α	0.581 (0.207)	1.141 (0.591)	-0.560 (0.562)	<0.0001
Decay test, β	0.896 (0.081)	0.972 (0.192)	-0.077 (0.201)	0.0011
Pair test, proportion epistatic of total	0.191 (0.093)	0.045 (0.080)	0.146 (0.122)	<0.0001
Pair test, proportion epistatic of non-lethal	0.743 (0.243)	0.781 (0.234)	-0.038 (0.303)	0.4374
Pair test, proportion synergistic of epistatic	0.271 (0.093)	0.168 (0.159)	0.103 (0.175)	<0.0001

* Two-tailed Wilcoxon signed-ranks test of the differences between 87 paired complex and simple organisms.

replicating computer programs that compete for central processing unit (CPU) time, which is the fuel needed for their replication. The programs mutate at random and evolve in a defined computational environment. Each digital organism has a genome length measured as the number of sequential instructions in its program. The sequence of instructions can change by mutation, including insertion and deletion events as well as point mutations that change one instruction to another. There are 28 different instructions, which can be thought of as analogous to the 20 different amino acids strung together in proteins. There is no imposed limit to the genome length of these digital organisms.

Starting from a short ancestral program, we generated 87 different ‘complex’ organisms by allowing replicated populations to evolve in an environment in which: (1) the baseline allocation of CPU time is proportional to genome size; and (2) certain mathematical operations, which require novel combinations of instructions, are rewarded with additional CPU time. For example, digital organisms may be rewarded for performing an ‘XOR’ operation (‘exclusive-or’ in which A or B is true, but not both) on 32-bit inputs using a series of ‘NAND’ operations (‘not-and’ in which A and B are not simultaneously true). ‘XOR’ is among the more complicated logical operations; ‘NAND’ is the only logical operator given as an instruction in Avida. In essence, these operations are a kind of metabolism that allows the digital organisms to acquire the CPU time needed for their replication. Starting from each complex organism, we derived a ‘simple’ organism that evolved in an environment that favoured faster replication and nothing else: (1) the allocation of CPU time is independent of genome length; and (2) mathematical operations are not rewarded. Thus, we define simple and complex organisms by the different environments in which they evolved. However, we cannot exclude the possibility that aspects other than functional complexity might contribute to differences between the two classes; for example, relaxed selection

for genome size in complex organisms may promote genetic redundancy, with attendant consequences for mutational effects.

All 87 simple digital organisms have shorter genomes than their paired complex progenitors, with mean lengths of 19.8 and 91.3 instructions, respectively (Table 1). In every case, both simple and complex digital organisms became more fit than their progenitors in their respective environments, where fitness is simply an organism's replication rate. Using all 174 complex and simple organisms, we then performed two experiments to investigate the effects of mutations and their interactions with respect to fitness. In the first experiment—the 'decay test'—we generated, for each organism, every possible mutant program that contained exactly one point mutation, plus a million (or more) programs with between two and ten random mutations. For each mutant, we measured fitness relative to its unmutated parent in the parent's selective environment. Then, for each parent, we regressed the average mutant fitness W versus mutation number M using a power function: $\log_{10} W = -\alpha M^\beta$. In biological terms, α describes the rate at which fitness decays under a multiplicative hypothesis, whereas β describes the form of 'epistasis'. If $\beta = 1$, mutational effects are multiplicative; if $\beta > 1$, each successive mutation tends to reduce fitness more than previous ones (synergistic epistasis); and if $\beta < 1$, each additional mutation is progressively less damaging on average (antagonistic epistasis). The parameter α is measured without error because $W = 1$ when $M = 0$ (by definition) and every possible mutant with $M = 1$ is tested, whereas β is estimated from sample data. The power function fits these data very well; unlike a quadratic function⁴, it never predicts that the decay curve bends upwards at high M .

Simple organisms are much more fragile than their complex counterparts with respect to the fitness effects of single mutations, as indicated by higher values of α ($P < 0.0001$; Table 1). This difference occurs because more mutations are lethal (prevent self-replication) in smaller genomes than in larger genomes (Fig. 1); on average, 92% and 53% of single mutations were lethal to simple and complex organisms, respectively. Partly offsetting this effect, non-lethal mutations are less damaging to the simple digital organisms; one non-lethal mutation reduces fitness by 11%, on average, in simple organisms, but by 44% in complex ones. The simple organisms do nothing except self-replicate; mutations that disrupt self-replication tend to be lethal, whereas most others are fairly harmless. The complex organisms perform mathematical functions in addition to self-replicating; most mutations that impact these functions hinder performance but are not lethal.

Simple and complex digital organisms also differ in terms of β , which indicates the average form of interaction ($P = 0.0011$; Table 1). In complex organisms, successive mutations tend to reduce fitness less than would be expected if effects were independent ($\beta < 1$), that is, complex organisms are also robust to the cumulative effect of multiple mutations. By contrast, successive mutations do not deviate significantly from multiplicative effects ($\beta = 1$) in simple organisms ($P = 0.2360$, Wilcoxon signed-ranks test). Figure 2 shows the difference between simple and complex digital organisms in their fitness decay curves, including the effects of both α and β .

The fact that simple organisms appear to show multiplicative effects of mutations on average fitness may mean either that there is little epistasis or that epistasis is widespread but different sets of mutations interact in opposite ways, obscuring the overall signal^{4,27}. We ran a second experiment—the 'pair test'—to distinguish between these two possibilities and gain further insight into the differences between simple and complex organisms. Instead of measuring fitness as progressively more mutations are added, as in the decay test, we examined numerous pairs of mutations by comparing the actual fitness of each double mutant with the expected fitness assuming multiplicative effects of the component mutations. The pair test shows that epistasis is more common in

complex organisms than in their simple counterparts ($P < 0.0001$; Table 1), with 19% of all mutation pairs deviating from multiplicative effects in the average complex organism (Fig. 3a) compared with <5% in the average simple organism (Fig. 3b). Frequencies of epistatic interactions are much higher if we exclude lethal pairs, in which actual and expected fitnesses are both zero (Fig. 3c, d). Excluding lethal pairs, which are much more common in simple organisms, there is no difference between classes in the prevalence of epistasis ($P = 0.4374$; Table 1). In both classes of digital organisms, epistatic interactions include a mixture of synergistic and antagonistic effects, and antagonistic effects are more common than synergy. Complex organisms are more prone to synergistic effects when expressed as a percentage of epistatic interactions ($P < 0.0001$; Table 1), but the overall excess of antagonism is greater in complex organisms. The failure of the decay test to find a significant deviation from multiplicative effects in the simple organisms evidently reflects a combination of two factors: epistasis is infrequent, and synergistic and antagonistic interactions oppose one another.

In summary, mutations in digital organisms frequently exhibit epistasis, including a diverse mixture of synergistic and antagonistic interactions. Such interactions are especially pronounced in the complex digital organisms that evolved large genomes and are rewarded for mathematical operations beyond self-replication. Frequent epistasis, including a mix of synergistic and antagonistic effects, also exists in a variety of real organisms^{4–6}. Thus, digital organisms experience complicated responses to genetic perturbations that appear similar to those seen in real organisms. The genomes and performances of digital organisms can be studied with far greater replication and precision than can be achieved with any real organism. Digital organisms may therefore offer a useful tool for addressing other biological questions in which complexity is both a barrier to understanding and an essential feature of the whole living system. Moreover, digital organisms allow us to test general hypotheses with a system that is built upon an artificial chemistry completely different from that used by real organisms. □

Methods

Evolution of digital organisms. Experiments were performed using version 1.3 of Avida, which can be obtained from <http://www.krl.caltech.edu/avida/pubs/nature99>. We used default settings unless otherwise indicated. Our first step was to generate pairs of complex and simple organisms. All evolution experiments began with the population at its carrying capacity (3,600 individuals). The probability of point mutation was 0.0075 per instruction copied; the probabilities of insertion and deletion mutations were each 0.05 per genome divide. Time is measured in arbitrary units called updates; every update represents the execution of an average of 30 instructions per individual in the population. (A typical generation is 5–10 updates, depending on genome size and execution.) Starting with the default ancestor (genome length 20), a population of complex organisms was propagated for 50,000 updates in an environment in which the baseline allocation of CPU time was proportional to genome size and extra CPU time was obtained by performing mathematical operations in the default task set. The former condition eliminates selection for smaller genomes, and the latter condition imposes selection for functional complexity. The rewards for performing operations are also specified in the default task set; they are scaled by their approximate difficulty and combined in a multiplicative fashion with one another and with the baseline CPU time. (Note that multiplicative scaling of phenotypic rewards does not imply multiplicative effects of mutations.) Using random number seeds, 87 complex populations were derived; subsequent experiments used the numerically dominant genotype from each population. Starting with each complex organism as progenitor, a population of simple organisms evolved for 25,000 updates by allocating the same CPU time to all organisms. There was selection for smaller genome size to promote faster replication and no selection for mathematical operations. Subsequent experiments used the most abundant genotype from each population.

Mutational analyses. We developed three genetic tools to analyse the effects of

point mutations on the performance of digital organisms. In all cases, the fitness (replication rate) of each mutant was calculated in the same environment in which its simple or complex parent evolved, and the mutant's fitness is expressed relative to the parent. The first tool makes every possible one-step point mutant for a particular parent. The default set includes 28 different instructions; given a parent of genome length 80, for example, there are $80 \times (28 - 1) = 2,160$ different one-step point mutants. The mean fitness of these mutants permits exact calculation of α in the decay test. The second tool produces a random sample of progeny that differ from their parent by two or more point mutations. For each parent, we generated between 10^5 and 10^7 progeny with two mutations, three mutations and so on, up to ten mutations. The third tool produces and analyses pairs of point mutations alone and in combination; for each two-step mutant, we have both corresponding one-step mutants. Having the single mutants allows us to compare a double mutant's actual fitness with the exact value expected under the hypothesis that the mutations interact in a multiplicative manner. We ran the pair test on 10^4 and 10^5 mutational pairs for each complex and simple organism, respectively.

Statistical methods. We performed the Wilcoxon signed-ranks test on the difference scores for all comparisons between complex and simple organisms²⁹. This test reflects the evolutionary relationship between pairs of organisms; it is also non-parametric and thus insensitive to deviations from a normal distribution. To estimate β in the decay tests, we minimized the sum of squared deviations around the log-transformed mean fitness values. We excluded samples with fewer than 100 viable mutants, in which case log mean fitness was poorly estimated. By increasing sample size to 10^8 , we can obtain additional viable mutants; the exclusion of some values because of insufficient sampling appears to have no systematic effect on estimation of β .

Received 12 April; accepted 27 May 1999.

- Ray, T. S. in *Artificial Life II* (eds Langton, C. G., Taylor, C., Farmer, J. D. & Rasmussen, S.) 371–408 (Addison-Wesley, Redwood City, California, 1991).
- Adami, C. Learning and complexity in genetic auto-adaptive systems. *Physica D* **80**, 154–170 (1995).
- Adami, C. *Introduction to Artificial Life* (Springer, New York, 1998).
- Elena, S. F. & Lenski, R. E. Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
- De Visser, J. A. G. M., Hoekstra, R. F. & van den Ende, H. Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. *Evolution* **51**, 1499–1505 (1997).
- Clark, A. G. & Wang, L. Epistasis in measured genotypes: *Drosophila* P-element insertions. *Genetics* **147**, 157–163 (1997).
- Wright, S. *Evolution and the Genetics of Populations* (Univ. Chicago Press, 1977).
- Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
- Kauffman, S. A. *The Origins of Order* (Oxford Univ. Press, New York, 1993).
- Paquin, C. & Adams, J. Relative fitness can decrease in evolving populations of *S. cerevisiae*. *Nature* **306**, 368–371 (1983).
- Dykhuizen, D. E., Dean, A. M. & Hartl, D. L. Metabolic flux and fitness. *Genetics* **115**, 25–31 (1987).
- Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl Acad. Sci. USA* **91**, 6808–6814 (1994).
- Rosenzweig, R. F., Sharp, R. R., Treeves, D. S. & Adams, J. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics* **137**, 903–917 (1994).
- Travisano, M., Mongold, J. A., Bennett, A. F. & Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **267**, 87–90 (1995).
- Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394**, 69–72 (1998).
- Burch, C. L. & Chao, L. Evolution by small steps and rugged landscapes in the RNA virus $\phi 6$. *Genetics* **151**, 921–927 (1999).
- De Visser, J. A., Zeyl, C. W., Gerrish, P. J., Blanchard, J. L. & Lenski, R. E. Diminishing returns from mutation supply rate in asexual populations. *Science* **283**, 404–406 (1999).
- Turner, P. E. & Chao, L. Prisoner's dilemma in an RNA virus. *Nature* **398**, 441–443 (1999).
- Maynard Smith, J. Byte-sized evolution. *Nature* **335**, 772–773 (1992).
- Holland, J. H. *Adaptation in Natural and Artificial Systems* (MIT Press, Cambridge, Massachusetts, 1992).
- Koza, J. R. *Genetic Programming* (MIT Press, Cambridge, Massachusetts, 1992).
- Frank, S. A. in *Adaptation* (eds Rose, M. R. & Lauder, G. V.) 451–505 (Academic, New York, 1996).
- Koza, J. R., Bennett, F. H., Andre, D. & Keane, M. A. in *Evolutionary Robotics* (ed. Gomi, T.) 37–76 (AAI Press, Kanata, Canada, 1998).
- Maynard Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, 1978).
- Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).
- Hurst, L. D. & Peck, J. R. Recent advances in understanding of the evolution and maintenance of sex. *Trends Ecol. Evol.* **11**, 46–52 (1996).
- Otto, S. P. & Feldman, M. W. Deleterious mutations, variable epistatic interactions, and the evolution of recombination. *Theor. Popul. Biol.* **51**, 134–147 (1997).
- Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
- Sokal, R. R. & Rohlf, F. J. *Biometry* 3rd edn (Freeman, New York, 1994).

Acknowledgements. We thank A. De Visser, S. Elena, D. Lenski, P. Moore, A. Moya and S. Remold for comments, discussion and technical assistance. Access to a Beowulf system was provided by the Center for Advanced Computing Research at the California Institute of Technology. This work was supported by an NSF grant to C.A. and a fellowship from the MacArthur Foundation to R.E.L.

Correspondence should be addressed to R.E.L. (e-mail: lenski@pilot.msu.edu); requests for materials to C.O. (e-mail: charles@krl.caltech.edu) or see <http://www.krl.caltech.edu/avida/pubs/nature> 99.

A general model for the structure and allometry of plant vascular systems

Geoffrey B. West^{*†}, James H. Brown^{†‡} & Brian J. Enquist^{†‡}

^{*} Theoretical Division, T-8, MS B285, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

[†] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

[‡] Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA

Vascular plants vary in size by about twelve orders of magnitude, and a single individual sequoia spans nearly this entire range as it grows from a seedling to a mature tree. Size influences nearly all of the structural, functional and ecological characteristics of organisms^{1,2}. Here we present an integrated model for the hydrodynamics, biomechanics and branching geometry of plants, based on the application of a general theory of resource distribution through hierarchical branching networks³ to the case of vascular plants. The model successfully predicts a fractal-like architecture and many known scaling laws, both between and within individual plants, including allometric exponents which are simple multiples of 1/4. We show that conducting tubes must taper

Box 1 Notation and geometry

The model can be described as a continuously branching hierarchical network running from the trunk (level 0) to the petioles (level N), with an arbitrary level denoted by k (Fig. 1). The architecture is characterized by three parameters (a , $\bar{a}$ and n), which relate daughter to parent branches: ratios of branch radii, $\beta_k = r_{k+1}/r_k = n^{-a/2}$, tube radii, $\tilde{\beta}_k = a_{k+1}/a_k = n^{-\bar{a}/2}$, and branch lengths, $\gamma_k = l_{k+1}/l_k$ and also the branching ratio, n , the number of daughter branches derived from a parent branch. Because the total number of tubes is preserved at each branching, $n = n_{k+1}/n_k$, where n_k is the number of tubes in a k th-level branch; n is taken to be independent of k and typically equals 2. Clearly, $n_k = n_N n^{N-k}$, where N is the total number of branching generations from trunk to petiole, and n_N is the number of tubes in a petiole, which is taken to be an invariant. Now, for a volume-filling network, $\gamma_k = n^{-1/3}$, independent of k (ref. 3). If tube tapering is uniform, $\bar{a}$ is also independent of k , and it follows that

$$\frac{r_k}{r_N} = n^{(N-k)a/2}; \quad \frac{a_k}{a_N} = \left(\frac{r_k}{r_N}\right)^{\bar{a}/a}; \quad \frac{l_k}{l_N} = \left(\frac{r_k}{r_N}\right)^{2/3a} \quad (1)$$

Various scaling laws can now be derived. For example, the number of terminal branches or leaves distal to the k th branch, $n_k^t = n_k/n_N = n^{N-k} = (r_k/r_N)^{2a}$, and the area of conductive tissue (CT), $A_k^{CT} = n_k \pi \bar{a}_k^2 = A_N^{CT} (r_k/r_N)^{2(1+\bar{a})/a}$, where $A_N^{CT} = n_N \pi \bar{a}_N^2$ is the area of conductive tissue in a petiole. Thus, the area of conductive tissue relative to the total (tot) branch cross-sectional area ($A_k^{\text{tot}} = \pi r_k^2$) is given by

$$f_k = \frac{A_k^{CT}}{A_k^{\text{tot}}} = n_N \left(\frac{\bar{a}_N}{r_N}\right)^2 \left(\frac{r_k}{r_N}\right)^{2(1+\bar{a})/a} \quad (2)$$

The total cross-sectional area scales as $n A_k^{\text{tot}}/A_k^{\text{tot}} = n \beta_k^2 = n^{1-a}$. When $a = 1$ this reduces to unity and the branching is area-preserving; that is, the cross-sectional area of the daughter branches is equal to that of the parent: $n A_{k+1}^{\text{tot}} = A_k^{\text{tot}}$. A simple example of this, considered in ref. 3, is the pipe model⁶, in which all tubes have the same constant diameter ($\bar{a} = 0$), are tightly bundled and have no non-conducting tissue. Here we consider the more realistic case in which tubes are loosely packed in sapwood and there may be non-conducting heartwood providing additional mechanical stability.

and, consequently, that the resistance and fluid flow per tube are independent of the total path length and plant size. This resolves the problem of resistance increasing with length, thereby allowing plants to evolve vertical architectures and explaining why the maximum height of trees is about 100 m. It also explains why the energy use of plants in ecosystems is size independent.

Most size-related variation can be characterized by allometric scaling laws of the form $Y = Y_0 M^b$, where Y is the variable of interest, Y_0 is a normalization constant, M is body mass and b is the scaling exponent. Although an enormous amount of anatomical and physiological data exist for plants^{2,4-17}, no single model has explained these diverse phenomena. We have modelled the transport of fluid from the trunk to the petioles through the xylem vessels of angiosperms (flowering plants). The model is based on a few general principles: (1) the branching network is volume filling; (2) the leaf and petiole size are invariant; (3) biomechanical constraints are uniform; and (4) energy dissipated in fluid flow is minimized. The model should also apply, with only minor modification, to transport through tracheids, phloem and roots.

The network is assumed to be composed of identical tubes of equal length running continuously in parallel from trunk to petiole (files of vessels connected in series). Tube diameter is taken to be constant within a branch segment but is allowed to vary between segments, thereby incorporating possible tapering. For simplicity, tapering within segments is ignored, as are the thickness and structure of the tube walls and connections between tubes. The notation and geometry of the model are described in Box 1, where it is shown that scaling relations can be parametrized in terms of just two exponents, a and $\bar{a}$, which determine how radii of branches and tubes scale within a plant. Box 2 shows how $\bar{a}$, and consequently the degree of tapering, are determined from hydrodynamic considerations, whereas a is determined from mechanical constraints.

The design of trunks and branches to resist buckling leads to some optimal relationship between their length and radius: $l_k \propto r_k^\alpha$. Comparison with equation (1) in Box 1 gives $a = 2/(3\alpha)$. If this holds uniformly for all branches, then α is constant, independent of k , in which case a and β_k are also constant. When coupled with the volume-filling constraint, $\gamma_k = n^{-1/3}$, this proves that the branching

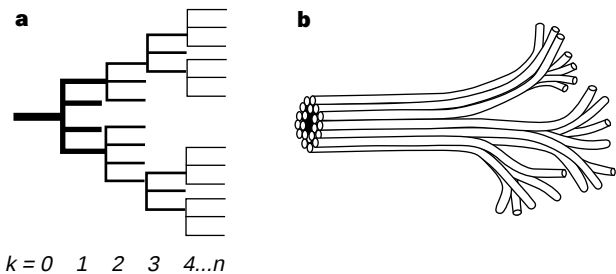


Figure 1 Branching structure. **a**, Topology of a plant branching network. **b**, Symbolic representation of branch vascular structure, showing conducting tubes and non-conducting tissue (black).

architecture is a self-similar fractal^{18,19}. Analyses based on solutions to the bending moment equations for beams give $\alpha = 2/3$ (refs 2, 20). This is most important for the trunk and large branches^{7,15,20}. Assuming that it holds for all branches (all k) leads to $a = 1$, which is the condition for area-preserving branching³. Unlike a previous report³, this derivation does not assume the pipe model. The result $a = 1$ implies that the leaf area distal to the k th branch $A_k^L = C_L r_k^2$, where $C_L \equiv a_L/r_N^2$ is invariant, a_L is the average area of a leaf, and r_N is the radius of a petiole. Taking $r_N \approx 0.5$ mm and $a_L \approx 30$ cm² gives $C_L \approx 1.2 \times 10^4$. The number of branches of a given size $N_k = n^k = n^N (r_N/r_k)^2 \propto r_k^{-2}$ (refs 2, 4, 6, 8, 12). These results are independent of $\bar{a}$.

The linear increase in hydrodynamic resistance with length implied by the classic Poiseuille formula⁴ for a uniform tube would preclude the very existence of tall trees, regardless of any mechanical constraint; for example, lower branches would be favoured over branches in the upper canopy where most light is collected. However, as shown in Box 2, if tubes taper, their total resistance from trunk to petiole Z_i need not increase with their total length. Indeed, for $\bar{a} > 1/6$, Z_i is a constant independent of total tube length and is the same for all plants, regardless of size. This ensures an equal supply to all leaves, both at different heights within a plant, and in plants of different sizes growing in similar environments. To avoid excess tapering, $\bar{a}$ should be the minimum possible value consistent with constant Z_i , namely $1/6$.

Box 3 shows how general scaling relationships for various physical quantities are derived in terms of a and $\bar{a}$. We now consider some consequences in the idealized case, where $a = 1$ (the area-preserving branching condition derived from biomechanical constraints) and $\bar{a} = 1/6$ (derived from minimizing tapering and hydrodynamic resistance).

Box 2 Hydrodynamics and vessel tapering

The resistance, Z_k , of a single tube (i) within any branch segment (k) is given by the Poiseuille formula⁴ ($Z_k = 8\eta l_k / \pi a_k^4$), where η is the fluid viscosity. The total resistance of a tube running from the trunk to a petiole through successive branch segments is $Z_i = \sum_{k=0}^N Z_k$. Using equation (1), this can be expressed as

$$Z_i = \sum_{k=0}^N Z_k = \left[\frac{1 - [(n^{1/3} - 1)l_T/l_N]^{(1-\bar{a})}}{1 - n^{(1/3-\bar{a})}} \right] Z_N \quad (3)$$

where $Z_N = 8\eta l_N / \pi a_N^4$ is the resistance of a tube in the petiole, and the total tube length $l_T = \sum_{k=0}^N l_k = l_0 / (1 - n^{-1/3})$. The crucial point is that, when $l_T \gg l_N$, the behaviour of Z_i depends critically on the degree of tapering, that is, whether $\bar{a}$ is less than, more than, or equal to $1/6$. First, consider $\bar{a} < 1/6$, which includes the pipe model; then equation (3) gives $Z_i \propto (l_T/l_N)^{(1-\bar{a})}$, so resistance increases with path length, l_T . However, if $\bar{a} > 1/6$, then $Z_i \approx Z_N / [1 - n^{(1/3-\bar{a})}]$, which can be shown to be its minimum value. This has the remarkable property that it is a constant, independent of total tube length, l_T ; because Z_N is an invariant, Z_i is also the same for all plants, regardless of size. Taking $\bar{a} \rightarrow 1/6$ to minimize tapering, yet keeping Z_i constant, gives $\bar{a} \approx 1/6 + (2N \ln n)^{-1}$. Thus, for large N (tall trees), $\bar{a} \approx 1/6$; for example, with $r_0 \approx 25$ cm and $N \approx 20$, the correction is only $\sim 1/30$. For small trees, however, whose height $h \ll e^{(6\bar{a}-1)^{-1}} l_N / (n^{1/3} - 1) \approx$ a few metres, the correction can be large, leading to calculable deviations from results derived using $\bar{a} = 1/6$. At this precise value, Z_i increases logarithmically with l_T , so $\bar{a}$ should implicitly be taken as infinitesimally larger than $1/6$.

Box 3 Scaling relations and allometry

Scaling relations within an individual plant are easily derived. Here we present some general formulae expressed in terms of a and $\bar{a}$. Because tubes are in parallel, the resistance of a branch segment is given by $Z_k = Z_k / n_k = 8\eta l_k / \pi n_k a_k^4$ and its conductivity by $K_k \equiv l_k / Z_k = K_N (r_k/r_N)^{2(1+2\bar{a})}$, where $K_N = \pi n_N a_N^4 / 8\eta$ is the conductivity of a petiole. Similarly, leaf-specific conductivity (the conductivity per unit leaf area), $L_k \equiv K_k / n_k^L a_L = L_N (r_k/r_N)^{4\bar{a}}$. Alternatively, these can be expressed as a function of the area of conducting tissue: $K_k \propto (A_k^{\text{CT}})^{(1+2\bar{a})/(1+\bar{a})}$ and $L_k \propto (A_k^{\text{CT}})^{(2\bar{a})/(1+\bar{a})}$ (refs 8-10).

Allometric relations can be derived by considering the total wood volume, V_W , which, for constant density, is proportional to the total mass, M . Thus, $V_W = \sum_{k=0}^N \pi n_k^L r_k^2 l_k = (\gamma \beta^2)^{-N} V_N / [1 - n \gamma \beta^2]$, where $\gamma = n^{-1/3}$, $\beta = n^{-a/2}$ and V_N is the volume of a petiole. Consequently, the total number of terminal branches, or leaves, should scale as $n_0^L = n^N \propto M^{3/(1+3\bar{a})}$. Similarly, from equation (1), $l_k \propto M^{(1-k/N)/(a+3)}$ and $r_k \propto M^{(1-k/N)(3a/2(a+3))}$, so the length and radius of a k th-level branch scale more slowly with M than those of the trunk, which scale as $l_0 \propto M^{1/(a+3)}$, and $r_0 \propto M^{3a/(2(a+3))}$.

First, the conductivity of a branch segment scales as $K_k/K_N \propto (r_k/r_N)^{8/3} \propto (A_k^{CT})^{8/7}$ and its leaf-specific conductivity as $L_k/L_N \propto (r_k/r_N)^{2/3} \propto (A_k^{CT})^{2/7}$. For a petiole with $a_N \approx 10 \mu\text{m}$ and $n_N \approx 200$, its conductivity $K_N \approx 7 \times 10^{-10} \text{ m}^4 \text{ s}^{-1} \text{ MPa}^{-1}$. For comparison, the pipe model, where $a = 1$ and $\bar{a} = 0$, incorrectly gives $K_k \propto r_k^2$ and $L_k \propto r_k^0$.

Second, the total number of terminal branches or leaves scales as $n_0^L \propto n^N \propto r_0^2 \propto M^{3/4}$. Similarly, the length and radius of the trunk scale as $l_0 \propto M^{1/4}$ and $r_0 \propto M^{3/8}$, respectively². The total height of a tree, h , is equivalent to the length of a tube running from trunk to leaf: $h = l_T \approx l_0/(1 - \gamma)$, so $h \propto M^{1/4}$. The number of branching generations, N , grows only logarithmically with mass, $N \propto \ln M$, and can be estimated from $N = 2\ln(r_0/r_N)/\ln n$. For a tree with trunk diameter 50 cm, petiole radius 0.5 mm and $n = 2$, this gives $N \approx 18$. From equation (1), the tube radius scales as $a_k/a_N = (r_k/r_N)^{1/6} \propto M^{(1-k/N)/16}$, so, in the trunk relative to the petiole, $a_0/a_N = n^{N/12} \propto M^{1/16}$. Taking $N \approx 18$ and $n = 2$, this gives $a_0/a_N \approx 2.8$, so, if $a_N \approx 10 \mu\text{m}$, then $a_0 \approx 30 \mu\text{m}$. Furthermore, over variation in mass of 12 orders of magnitude, the radius of a xylem vessel in the trunk, a_0 , should change by only about 60% (refs 4, 11).

Third, the pressure gradient across an arbitrary branch, $\Delta P_k/l_k$ scales as $l_k^{-6\bar{a}} = l_k^{-1}$ and so is steeper in smaller branches than larger ones. In particular, the ratio between trunk and petiole is predicted to be $(\Delta P_0/l_0)/(\Delta P_N/l_N) = l_N/l_0 \propto M^{-1/4}$, and so is smaller in taller trees. If $l_0 \approx 4 \text{ m}$ and $l_N \approx 4 \text{ cm}$, then this ratio is $\sim 1/100$ (refs 4, 5). The flow rate through a single tube $\dot{Q}_i = \Delta P/Z_i$, where ΔP is the overall pressure difference between air and soil. Because Z_i and ΔP are both independent of plant size, $\dot{Q}_i$ must also be size independent; the flow rate in a tube in a large tree is therefore the same as that in a small plant. This is the origin of the 'energy equivalence relationship'²¹. Metabolic rate, B , the rate of gross photosynthesis, is proportional to the total volume flow rate through all n_0 vascular tubes, and is therefore given by $\dot{Q}_0 = n_0 \dot{Q}_i = n^N n_N \dot{Q}_i \propto M^{3/4}$ (refs 2, 21). Thus $B \propto M^{3/4}$.

Fourth, tubes taper, so fluid velocity, u_k , must increase in small branches. The flow rate through a single tube is $\dot{Q}_i = \pi a_k^2 u_k$, so $u_k \propto a_k^{-2} \propto r_k^{-1/3}$. Thus, the ratio of petiole to trunk velocities $u_N/u_0 = (r_0/r_N)^{1/3}$. Taking $r_N \approx 0.5 \text{ mm}$ and $r_0 \approx 50 \text{ cm}$ gives $u_N/u_0 \approx 4.6$. Allometrically, u_k scales as $M^{-(1-k/N)/8}$ so, for the trunk, $u_0 \propto M^{-1/8}$. Thus, over eight orders of magnitude variation in mass (roughly a 50-cm sapling relative to a 50-m tree), fluid velocity in the trunk is predicted to decrease by a factor of ~ 10 (ref. 4).

Fifth, a particularly sensitive test of the model is its accurate prediction of how total plant resistance changes as progressively larger branch segments are removed (Box 4).

Finally, if tapering continued indefinitely, vessel radii in the trunk would become too large, leading to a maximum height for trees. In Box 5 we show that $h^{\text{max}} \sim 100 \text{ m}$.

Although the model makes several simplifying assumptions, its power rests on fundamental physical and biological principles as well as realistic features of the architecture, biomechanics and hydrodynamics of vascular plants. It accurately predicts scaling exponents (Table 1), their normalizations and the magnitude of

Box 4 Removing branch segments

Suppose that only branches up to the k th level remain. Because the total number of tubes in the trunk is still n_0 , the total resistance of the remaining branch network is $Z_k^{\text{TOT}} = (\Sigma_{i=0}^k Z_i^*)/n_0$; note that Z_N^{TOT} is the total resistance of the uncut tree. The ratio of these, $R_k = Z_k^{\text{TOT}}/Z_N^{\text{TOT}}$, is given by $R_k = [(r_{k+1}/r_0)^p - 1]/[(r_{N+1}/r_0)^p - 1]$, where $p = 2(1 - 6\bar{a})/3a$ and $r_{N+1} = r_N n^{-a/2}$. Note that $R_N = 1$ and $R_{-1} = 0$, the latter corresponding to the limit where all conducting tissue has been removed. As shown in Fig. 2, our prediction, $\bar{a} = 1/6$, agrees very well with measured values¹² and provides a much better fit than the classic pipe model, $\bar{a} = 0$.

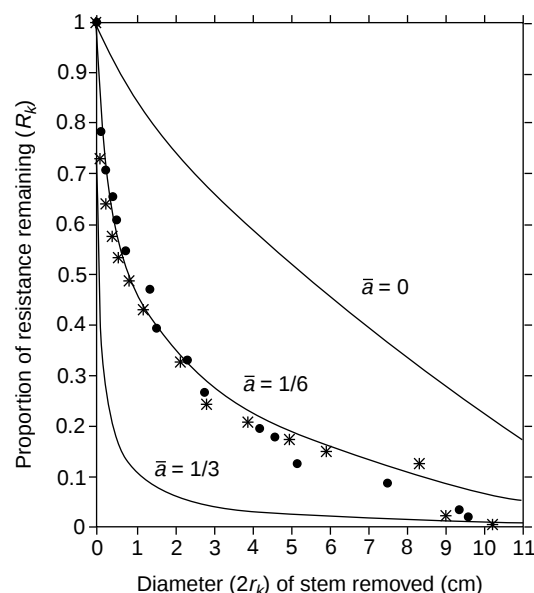


Figure 2 Effect of sequentially removing branches of increasing radius, r_k , on the proportion of total resistance remaining, R_k . The predicted value, $\bar{a} = 1/6$, is in excellent agreement with the data¹² that represent two trees. In contrast, $\bar{a} = 0$, corresponding to no tapering of tubes as in the pipe model⁶, and $\bar{a} = 1/3$, give poor agreement. The curves terminate when only the trunk remains. If extrapolated, all would converge at the trunk diameter, $\sim 14.5 \text{ cm}$, where $r_k = r_0$ and no conducting tissue remains, such that $R_k = R_1 = 0$.

many variables such as leaf area supplied by a tube and conductivity of a branch segment^{8,10–12}. We are unaware of any data that seriously disagree with its predictions. Because this zeroth-order model describes an 'average idealized' plant, it can serve as a starting point for more elaborate models that incorporate special features of plants of different taxa or growing in different environments. Such complications include: (1) variation in vessel size within and among plants growing in different environments; (2) horizontal flow between parallel tubes; (3) variation in tube length and branching symmetry (but see ref. 22); (4) variation in α , that is, not all branches are subject to the same biomechanical constraint^{2,15}; (5) departures from precise volume filling in plants such as palms, vines,

Box 5 Area of conducting tissue and the maximum height of trees

From the argument leading to equation (2), we have $A_k^{CT} \propto r_k^{2(1+\bar{a})/a} \propto A_N^{TOT7/6}$. This can be expressed in terms of the number of leaves, n_k^L , distal to the k th level branch: $A_k^{CT} \propto (n_k^L)^{1+\bar{a}} \propto (n_k^L)^{7/6}$. Now, from equation (2), the proportion of conductive tissue relative to total cross-sectional area of a branch scales as $f_k \propto r_k^{2(1+\bar{a})-a/a} \propto r_k^{1/3}$ (refs 8, 13, 14). For the trunk, in particular, we necessarily have $f_0 \leq 1$, so equation (2) leads to a limitation on the radius and consequently the height of a tree:

$$h^{\text{max}} = \frac{l_N}{(1 - n^{-1/3})} \left(\frac{r_N^2}{a_N^2 n_N} \right)^{1/(1+\bar{a}-a)} ; \quad r_0^{\text{max}} = r_N \left(\frac{r_N^2}{a_N^2 n_N} \right)^{a/(2(1+\bar{a}-a))}$$

With $a = 1$ and $\bar{a} = 1/6$, $h^{\text{max}} = l_N n_N^4 / [a_N^4 n_N^2 (1 - n^{-1/3})]$ and $r_0^{\text{max}} = r_N^7 / a_N^6 n_N^3$. Because of the large exponents, these formulae are very sensitive to the parameters of the petiole. For example, take $r_N = 0.5 \text{ mm}$, $a_N = 10 \mu\text{m}$, and $n_N = 200$, then $r_0^{\text{max}} \approx 1 \text{ m}$ and $h^{\text{max}} \approx 40 \text{ m}$; if, instead, $a_N = 8 \mu\text{m}$, then $r_0^{\text{max}} \approx 4 \text{ m}$ and $h^{\text{max}} \approx 100 \text{ m}$. Thus, although these formulae cannot be used to accurately estimate maximum size, they do show why h^{max} is of the order of 100 m rather than 1 m or 1,000 m. This provides a fundamental mechanical and hydrodynamic explanation why the size of trees is limited. If the network were not optimized, then h^{max} would be substantially reduced; for example, if $\bar{a} = 1/3$, then $h^{\text{max}} \approx 1 \text{ m}$. Thus, if trees are to grow tall, $\bar{a}$ must approach the optimal minimum value, $1/6$.

Table 1 Predicted values of scaling exponents for physiological and anatomical variables of plant vascular systems.

Variable	Plant mass		Branch radius		
	Exponent predicted	Symbol	Symbol	Exponent	
				Predicted	Observed
Number of leaves	$\frac{3}{4}$ (0.75)	n_0^L	n_k^L	2 (2.00)	2.007 (ref. 12)
Number of branches	$\frac{3}{4}$ (0.75)	N_0	N_k	-2 (-2.00)	-2.00 (ref. 6)
Number of tubes	$\frac{3}{4}$ (0.75)	n_0	n_k	2 (2.00)	n.d.
Branch length	$\frac{1}{4}$ (0.25)	l_0	l_k	$\frac{2}{3}$ (0.67)	0.652 (ref. 6)
Branch radius	$\frac{3}{8}$ (0.375)	r_0			
Area of conductive tissue	$\frac{7}{8}$ (0.875)	A_0^{CT}	A_k^{CT}	$\frac{7}{3}$ (2.33)	2.13 (ref. 8)
Tube radius	$\frac{1}{16}$ (0.0625)	a_0	a_k	$\frac{1}{3}$ (0.167)	n.d.
Conductivity	1 (1.00)	K_0	K_k	$\frac{5}{3}$ (2.67)	2.63 (ref. 12)
Leaf-specific conductivity	$\frac{1}{4}$ (0.25)	L_0	L_k	$\frac{2}{3}$ (0.67)	0.727 (ref. 17)
Fluid flow rate			$\dot{Q}_k$	2 (2.00)	n.d.
Metabolic rate	$\frac{3}{4}$ (0.75)	$\dot{Q}_0$			
Pressure gradient	$-\frac{1}{4}$ (-0.25)	$\Delta P_0/l_0$	$\Delta P_k/l_k$	$-\frac{2}{3}$ (-0.67)	n.d.
Fluid velocity	$-\frac{1}{8}$ (-0.125)	u_0	u_k	$-\frac{1}{3}$ (-0.33)	n.d.
Branch resistance	$-\frac{3}{8}$ (-0.375)	Z_0	Z_k	$-\frac{1}{3}$ (-0.33)	n.d.
Tree height	$\frac{1}{4}$ (0.25)	h			
Reproductive biomass	$\frac{3}{4}$ (0.75)				
Total fluid volume	$\frac{25}{24}$ (1.0415)				

Values are given as a function of total plant mass, M , and branch radius, r_k . For the latter case, predictions are compared with measured values in the last column. References cited do not quote confidence levels, except for branch length, where they are given as ± 0.036 . Because botanists rarely report allometric scaling with mass, no values for observed exponents are quoted. n.d., no data available.

ferns, grasses and saplings with few branches, so that $\gamma \rightarrow n^{-1/2}$ rather than $n^{-1/3}$, leading to $l_k \propto r_k$ (refs 7, 15); and (6) constrictions in tubes at petioles and perhaps at other branch junctions^{4,16}. These complications are expected to have small effects, because many quantities, such as scaling exponents, effectively average out over the whole plant.

The model quantitatively predicts how vessels must taper to compensate for variation in total transport path length. This is supported by measured changes in vessel radius and resistance within and between plants^{4,11} and leads to a maximum height for trees. An important consequence is that tapering ensures comparable xylem flow to all leaves. Competition for light has apparently led to a design that maximizes canopy height and simultaneously minimizes tapering of vascular tubes. In a given environment with a fixed pressure differential between air and soil, on average all xylem tubes of all plants conduct water and nutrients at approximately the same rate. This counterintuitive result provides the fundamental basis for the recently demonstrated equivalence of resource use, independent of plant size, across diverse ecosystems²¹.

The model shows that quarter-power allometric scaling laws, which are well known in animals¹, also apply to many characteristics of plants². There are many parallels: in both, metabolic rate scales as $M^{3/4}$, radius of trunk and aorta as $M^{3/8}$, and size of and fluid velocity in terminal vessels as M^0 . It seems that these scaling laws are nearly universal in biology, and that they have their origins in common geometric and hydrodynamic principles that govern the transport of essential materials to support cellular metabolism. □

Received 27 November 1998; accepted 20 May 1999.

- Schmidt-Nielsen, K. *Scaling: Why is Animal Size so Important?* (Cambridge Univ. Press, Cambridge, 1984).
- Niklas, K. J. *Plant Allometry: The Scaling of Form and Process* (Univ. of Chicago Press, Chicago, 1994).
- West, G. B., Brown, J. H. & Enquist, B. J. A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126 (1997).
- Zimmermann, M. H. *Xylem Structure and the Ascent of Sap* (Springer, New York, 1983).
- Tyree, M. T. & Ewers, F. W. The hydraulic architecture of trees and other woody plants. *New Phytol.* **119**, 345–360 (1991).

- Shinozaki, K., Yoda, K., Hozumi, K. & Kira, T. A quantitative analysis of plant form—the pipe model theory. I. Basic analysis. *Jpn. J. Ecol.* **14**, 97–105 (1964).
- Bertram, J. E. A. Size-dependent differential scaling in branches: the mechanical design of trees revisited. *Trees* **4**, 241–253 (1989).
- Patino, S., Tyree, M. T. & Herre, E. A. Comparison of hydraulic architecture of woody plants of differing phylogeny and growth form with special reference to free standing and hemi-epiphytic *Ficus* species from Panama. *New Phytol.* **129**, 125–134 (1995).
- Kuuluvainen, T., Sprugel, D. G. & Brooks, J. R. Hydraulic architecture and structure of *Abies lasiocarpa* seedlings in three alpine meadows of different moisture status in the eastern Olympic mountains, Washington, U.S.A. *Arct. Alp. Res.* **28**, 60–64 (1996).
- Yang, S. & Tyree, M. T. Hydraulic architecture of *Acer saccharum* and *A. rubrum*: comparison of branches to whole trees and the contribution of leaves to hydraulic resistance. *J. Exp. Bot.* **45**, 179–186 (1994).
- Ewers, F. W. & Zimmermann, M. H. The hydraulic architecture of eastern hemlock *Tsuga canadensis*. *Can. J. Bot.* **62**, 940–946 (1984).
- Yang, S. & Tyree, M. T. Hydraulic resistance in *Acer saccharum* shoots and its influence on leaf water potential and transpiration. *Tree Physiol.* **12**, 231–242 (1993).
- Long, J. N., Smith, F. W. & Scott, D. R. M. The role of Douglas fir stem sapwood and heartwood in the mechanical and physiological support of crowns and development of stem form. *Can. J. For. Res.* **11**, 459–464 (1981).
- Rogers, R. & Hinckley, T. M. Foliar weight and area related to current sapwood area in oak. *Forest Sci.* **25**, 298–303 (1979).
- Niklas, K. J. Size-dependent allometry of tree height, diameter and trunk taper. *Ann. Bot.* **75**, 217–227 (1995).
- Tyree, M. T. & Alexander, J. D. Hydraulic conductivity of branch junctions in three temperate tree species. *Trees* **7**, 156–159 (1993).
- Tyree, M. T., Graham, M. E. D., Cooper, K. E. & Bazos, L. J. The hydraulic architecture of *Thuja occidentalis*. *Can. J. Bot.* **61**, 2105–2111 (1983).
- Fitter, A. H. & Strickland, T. R. Fractal characterization of root system architecture. *Funct. Ecol.* **6**, 632–635 (1992).
- Morse, D. R., Lawton, J. H., Dodson, J. H. & Williamson, M. M. Fractal dimension of vegetation and the distribution of arthropod body lengths. *Nature* **314**, 731–733 (1985).
- McMahon, T. A. & Kronauer, R. E. Tree structures: deducing the principle of mechanical design. *J. Theor. Biol.* **59**, 443–436 (1976).
- Enquist, B. J., Brown, J. H. & West, G. B. Allometric scaling of plant energetics and population density. *Nature* **395**, 163–165 (1998).
- Turcotte, D. L., Pelletier, J. D. & Newman, W. I. Networks with side branching in biology. *J. Theor. Biol.* **193**, 577–592 (1998).

Acknowledgements. We thank K. Niklas and M. Tyree for comments. J.H.B. was supported by the NSF, B.J.E. by the NSF and with an NSF post-doctoral fellowship, and G.B.W. by the US Department of Energy. We also thank the Thaw Charitable Trust for support.

Correspondence and requests for materials should be addressed to B.J.E. (e-mail: benquist@unm.edu).

Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*

Eli A. Stahl*, Greg Dwyer†, Rodney Mauricio†, Martin Kreitman*† & Joy Bergelson*†

*Committee on Genetics, †Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

The co-evolutionary ‘arms race’¹ is a widely accepted model for the evolution of host–pathogen interactions. This model predicts that variation for disease resistance will be transient, and that host populations generally will be monomorphic at disease-resistance (*R*-gene) loci. However, plant populations show considerable polymorphism at *R*-gene loci involved in pathogen recognition². Here we have tested the arms-race model in *Arabidopsis thaliana* by analysing sequences flanking *Rpm1*, a gene conferring the ability to recognize *Pseudomonas* pathogens carrying *AvrRpm1* or *AvrB* (ref. 3). We reject the arms-race hypothesis: resistance and susceptibility alleles at this locus have co-existed for millions of years. To account for the age of alleles and the relative levels of polymorphism within allelic classes, we use coalescence theory to model the long-term accumulation of nucleotide polymorphism in the context of the short-term ecological dynamics of disease resistance. This analysis supports a ‘trench warfare’ hypothesis, in which advances and retreats of resistance-allele frequency maintain variation for disease resistance as a dynamic polymorphism^{4,5}.

Arabidopsis thaliana exhibits a disease-resistance polymorphism in which susceptible individuals are completely lacking *Rpm1* (refs 3, 6). The presence of *Rpm1* homologues in *Brassica*⁶ and the closely

We used population genetic theory to formulate testable predictions arising from the arms-race hypothesis. Specifically, continual selective turnover of alleles is a hallmark of this model⁹, and theory for selective sweeps¹⁰ predicts that resistance alleles should be young and nearly identical in sequence to susceptibility alleles. In contrast, long-lived variation for resistance is an alternative outcome of ecological models of gene-for-gene interactions^{11,12}, and theory for balanced polymorphism¹³ predicts old resistance and susceptibility alleles that differ substantially in DNA sequence, especially near the site under selection. These alternative hypotheses make opposing predictions about variation among alleles, and they can be distinguished by testing data against a null model of neutral molecular evolution¹⁴.

Genotypes are presence or absence of *Rpm1* determined by PCR. Both alleles of Kz-13 were included in population genetic analyses on the assumption that heterozygosity resulted from recent outcrossing, so the 26 accession random sample includes 27 alleles. The observed resistance-allele frequency 0.52 is not sensitive to accessions' source, field collected (asterisks) versus stock centre, or geographic origin. Phenotypes after inoculation with *Pseudomonas syringae* DC3000 with *AvrRpm1* (see Methods) are resistance, disease or not determined (n.d.). All lines (except NfC-10, Tamm-07 and RF-4, which were not tested) exhibited disease when challenged with the control strain, DC3000 without *AvrRpm1*.

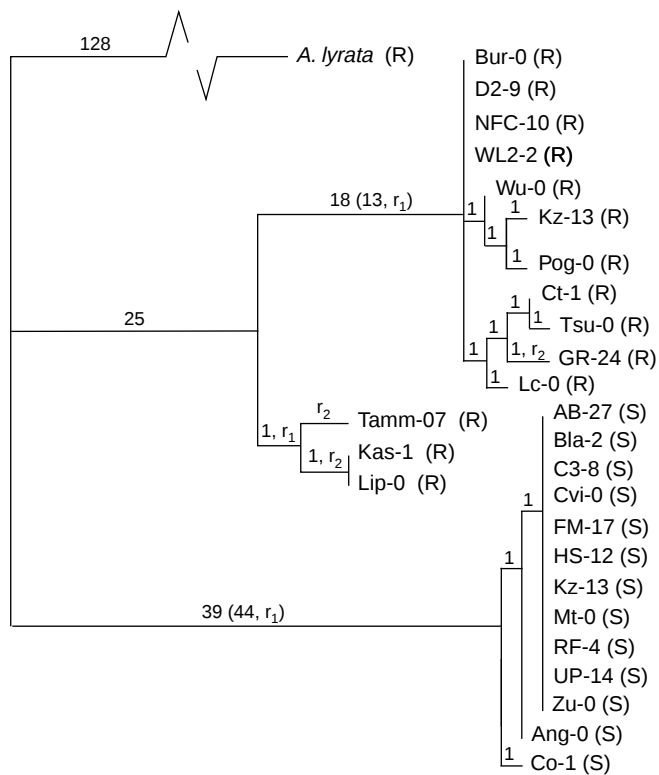


Figure 1 Genealogy of the *Rpm1* junction region. Alleles are named by accession, with *Rpm1* genotypes resistance (R) or susceptibility (S) in parentheses. This tree is one of fifteen most parsimonious trees, which differ in the branching within the R allelic class. The divergence between two clusters of R alleles is attributable to a recombination event (r_1) between R and S alleles at the 5' end of the sequenced region, involving 18 polymorphic sites (13 and 5 placed on the R and S lineages, respectively). Numbers above branches reflect a placement of nucleotide mutations taking into account two inferred recombination events, r_1 (above) and r_2 within the R allelic class. In assuming that mutations shared by non-recombinant R alleles (Bur-0 cluster) are differences between R and S alleles rather than within the R allelic class, this placement affects only the results of our intra-class polymorphisms analysis and is conservative.

sides. The ratio of polymorphism to divergence is significantly heterogenous across the junction region compared to neutral expectations (Runs test¹⁶, $P < 0.005$), owing to the peak of polymorphism associated with *Rpm1*. At the junction, divergence between resistance and susceptibility alleles reaches 11.8%, approximately equal to the 10.5% average divergence between *A. thaliana* and *A. lyrata*, and similar to species divergence for synonymous and noncoding sites at other loci^{17–20}. At a substitution rate of 6×10^{-9} per site per year²¹, this level of divergence indicates that the *Rpm1* polymorphism arose around 9.8 million years (Myr) ago.

Having rejected neutrality based on the Tajima's D and Runs tests, we can entertain two alternatives to natural selection maintaining resistance and susceptibility alleles. One possibility is that many nucleotide changes occurred in the same mutational event as the *Rpm1* deletion. If this were true then the point mutations should be associated with the susceptibility allele. However, there are similar numbers of changes on the resistance and susceptibility lineages (Fig. 1, $\chi^2_{\text{dif}} = 0.44$). Even if many point mutations occurred on both alleles in the individual suffering the *Rpm1* deletion of one allele, it is highly improbable that this particular undeleted resistance allele would have come to replace all other resistance alleles in the species.

A second alternative is that resistance and susceptibility alleles originated 9.8 Myr ago in a geographically subdivided population and have persisted in separate subpopulations until the recent past. Under this scenario, a signature of haplotypic divergence should be

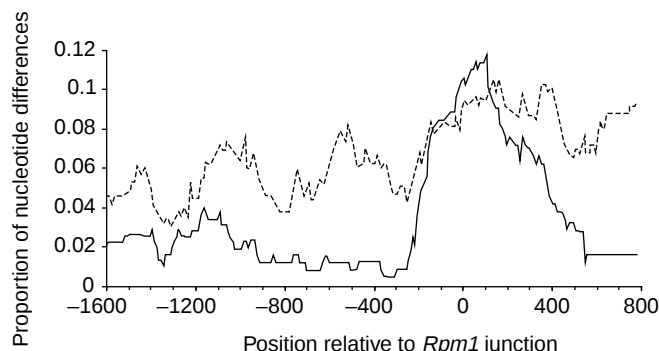


Figure 2 Sliding window analysis. The average pairwise proportion of nucleotide differences are shown between resistance and susceptibility alleles (solid line), and between all *A. thaliana* alleles and the *A. lyrata* sequence (dashed line). Values are midpoints of 250-bp windows (adjusted to exclude gaps). Position is relative to the *Rpm1* deletion (0 bp, not shown), so that negative and positive positions are, respectively, 5' and 3' of the *Rpm1* cds.

present across the genome. Our own coalescence simulations of subdivided populations (see Methods) find that the configuration of polymorphism at *Rpm1* is inconsistent with a subdivision explanation. Although the haplotypes corresponding to the two *Rpm1* allelic classes extend across the sequenced region, polymorphic sites are too clustered within the region. Thus, the *Rpm1* sequence data reject the arms-race hypothesis in favour of an alternative that leads to the selective maintenance of variation for disease resistance.

Although the arms-race hypothesis cannot explain the divergence between resistance and susceptibility alleles, a succession of positively selected resistance alleles is still possible. We examined divergence between *A. lyrata* and resistant *A. thaliana* accession Col-0 (ref. 3) across the entire *Rpm1* coding sequence to look for an accelerated amino-acid substitution rate indicative of positive selection¹⁴. On the contrary, amino-acid replacements have accumulated at a much slower rate than synonymous changes²² ($K_a = 0.028$ and $K_s = 0.13$). In addition, the 55 amino-acid substitutions between the two species are scattered throughout the coding sequence rather than being associated with the leucine-rich repeat region, the putative pathogen-recognition domain²³ ($\chi^2_{\text{ldf}} = 0.11$). Selectively driven turnover of resistance alleles does not appear to be important in RPM1 protein evolution.

Our data indicate, therefore, that the two alleles at *Rpm1* represent a long-lived polymorphism maintained by natural selection rather than the result of a co-evolutionary arms race. How does selection act to maintain this polymorphism? Because an *Rpm1*-null allele confers susceptibility, ecological trade-offs for allelic resistance specificities cannot apply here. Heterozygote advantage is also ruled out by *A. thaliana*'s extremely low outcrossing rate²⁵. Maintenance of this polymorphism therefore requires a balance between selection for disease resistance and a cost of resistance in the absence of the pathogen²⁴, and further requires temporal or geographic variation in selection.

Indeed, variability within allelic classes at *Rpm1* suggests temporal variation in selection. For a long-lived polymorphism maintained by natural selection at a constant frequency, the number of linked neutral segregating sites within an allelic class is expected to be proportional to its frequency¹³. However, resistance alleles of *Rpm1* segregate for ten times as many polymorphisms as do susceptibility alleles (30 versus 3), despite the fact that both alleles are approximately equally frequent. Excluding intraclass polymorphisms that may have arisen by rare recombination events between allelic classes (Fig. 1), we conservatively estimate that resistance and susceptibility allelic classes have 12 and 3 intraclass polymorphisms, respectively. Under a model of balancing selection

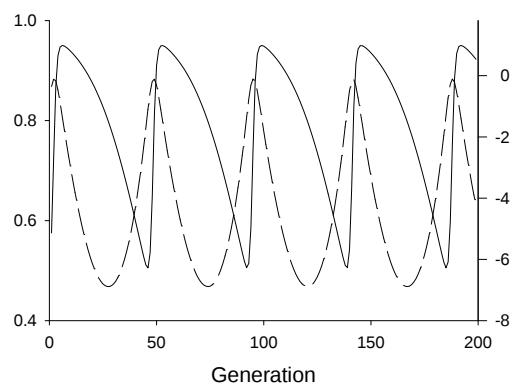


Figure 3 Model for an *A. thaliana*-*Pseudomonas* interaction. Host resistance allele frequency (solid line) and pathogen density (dashed line) are shown for a model that includes a cost of resistance and yearly frequency-dependent epidemics. The trajectory shown here gives a ratio of polymorphism within resistance and susceptibility allelic classes of 5.71, which compares well with the observed value of 5.72. Long-period cycles in the frequency of resistance are required for the model to achieve a good fit to the polymorphism data.

at a constant frequency (see Methods), these numbers of intraclass polymorphisms are expected for a resistance-allele frequency of 0.81, a significantly higher value than our worldwide estimate of 0.52. Furthermore, the resistance allelic class contains too many of the intraclass polymorphisms for a parametric *Rpm1* allele frequency of 0.52 ($P = 0.0377$). The data therefore indicate that the frequency of resistance at *Rpm1* may have fluctuated over time.

Temporal variation arises in host-pathogen interactions because disease spread is more likely when the frequency of resistance is low. This frequency-dependent selection can lead to the periodic recurrence of severe epidemics, causing the frequency of resistance to cycle^{11,12}. To determine whether frequency-dependent epidemics can account for the *Rpm1* population genetic data, we constructed and analysed a model describing an *A. thaliana*-*Pseudomonas* interaction. Our model differs from previous models in allowing *Pseudomonas* to survive between epidemics in nonpathogenic populations²⁵. Figure 3 shows model results leading to long-lived polymorphism and long-period cycles in the frequency of resistance. The cycles attain the observed resistance-allele frequency and give rise to the observed variability within allelic classes. Our data are consistent with a model in which variation for disease resistance is maintained by frequency-dependent selection, and thus provide, to our knowledge, the first empirical evidence for dynamical polymorphism, a phenomenon that Hamilton and co-workers^{5,12} have argued is likely to be widespread and important in the evolution of sex. Because in this model epidemics alternate with periods of high host resistance, leading to ceaseless advances and retreats for both host and pathogen, we use the term 'trench warfare' rather than 'arms race' to describe the evolution of this host-pathogen interaction.

Resistance-allele frequencies in local populations throughout the northern hemisphere are generally high (Fig. 4), as required for our model to accommodate the sequence data. A striking exception is the midwest United States, where resistance-allele frequencies are much lower than our model predicts. This indicates that the pathogen recognized by *Rpm1* may not be prevalent in the midwest.

We note that our model is deterministic, and in finite populations drift and founder effects can lead to the rapid loss of variation in local populations, whereas geographic variation can help to maintain polymorphism in the species as a whole²⁶. Dynamical polymorphism is the hallmark of the trench-warfare hypothesis, but, considering the effects of drift, trench warfare might not always lead to the long-term maintenance of variation for disease resistance.

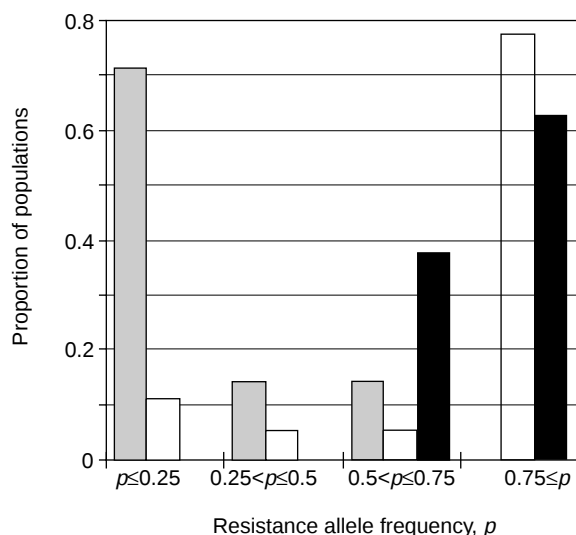


Figure 4 Resistance-allele frequency distributions. The proportion of populations with *Rpm1* resistance-allele frequencies in the indicated range for the United States midwest (stippled bars, 21 populations) and for the northern hemisphere excluding the US midwest (white bars, 19 populations) are compared with that predicted by our model for an *A. thaliana*–*Pseudomonas* interaction (black bars). The predicted distribution is the proportion of time during each cycle that the resistance allele spends in each frequency category, assuming that each population represents an independent random time point under the model. About ten individuals were sampled from each population.

Thus, the finding of a recently evolved disease resistance polymorphism in *Arabidopsis*, *Rps2* (ref. 27), is not inconsistent with the trench-warfare hypothesis. Population genetic studies of additional disease-resistance loci will be needed to evaluate the generality of this trench warfare model. Information on the ecological dynamics and demography of disease resistance provided by these studies should be valuable in applied contexts. □

Methods

Plant materials and phenotyping. Seeds of 26 accessions (Table 1) and *A. lyrata* were taken from the collections of J.B. and R.M., or received from the *Arabidopsis* stock centres. Accessions were phenotyped for *Rpm1* disease resistance by dipping plants in liquid cultures of *Pseudomonas syringae* DC3000 (control) or DC3000::AvrRpm1 (from J. Dangel), noting resistance response or disease symptoms after four days⁸. DNA from 447 individuals representing 40 populations of *A. thaliana* were assembled from the collections of J.B., R.M. and N. Miyashita. Seed stocks have been donated to the ABRC. Seeds were germinated in the greenhouse, and DNA was extracted according to standard protocols²⁸.

Genotyping. The *Rpm1* junction region sequence (J. Dangel and M. Grant, unpublished data) was extended to 2 kilobases (kb) 5' and 1 kb 3' of the junction, by inverse PCR, cloning and sequencing of accessions La-er and Nd-0. We used these sequences to design primers. A three-primer PCR (primers 1190⁺ TCGTGATTCCATTGCTTGTA, R1[−] TCGGGTGTGTTTCGTCAT and S1[−] GTGAGCAAAAGAGATGTG, designed by G. Wichman) gives alternative products for resistance and susceptibility alleles. Before cloning and/or sequencing the resistance alleles, the junction region was amplified using 5' primers (R5J⁺ CCCCAGAGGCTGCTGAGAAATGGC and R5I[−] CCATCATCAATAGGCGGTAATAATGCTGC) and 3' primers (R3I⁺ CGGTCAAGGGTTAAACACATTCCTGC and R3J[−] AACGGATCATCATTATTTC-AAGTACATCACATCGC). Susceptibility alleles were amplified using R5J⁺ and R3J[−]. An *A. lyrata* allele, including the junction region and the intronless *Rpm1* coding sequence, was assembled by shotgun-cloning and sequencing a long-PCR product (primers R5J⁺ and R3J[−]). Cloning was into a modified pZero 2.1 vector (Stratagene), and sequences were generated from PCR products of multiple colonies. The flanking sequences and the *A. lyrata* coding sequence have Genbank accession numbers AF122981–AF123027. Aligned

flanking sequences are available from the author.

Population genetics analyses. We performed Tajima's *D* test¹⁵, sliding window analyses and silent and replacement divergence²² estimation using DNAsp²⁹. Runs tests used the DNAslider program¹⁶, with recombination parameters from 0 to 32 (10,000 replicates for conservative recombination values). Population structure in *A. thaliana* was studied under a two-sub-population equilibrium migration model³⁰, conditional on the numbers of segregating sites, using simulation code from R. Hudson. We chose the migration parameter to maximize the probability of the homozygosity of derived mutations in several published datasets^{17–20}.

Balancing selection at a constant frequency was studied over a range of frequencies by simulation (10,000 replicates), assuming that samples of 14 resistance and 13 susceptibility alleles coalesce independently at rates proportional to their relative frequencies¹³ without recombination, and that the sum of the number of intraclass polymorphisms for the two allelic classes equals the observed sum of 15. We calculated the mean number of polymorphisms and the probability of 12 or more polymorphisms within the resistance allelic class.

Ecological modelling. In the absence of field observations of interactions between *A. thaliana* and *Pseudomonas*, we based our model on the biology of each species. Because *rpm1* is a deletion, it cannot encode specificity for an alternative pathogen. We assume that *P. syringae* strains that interact with *Rpm1* are specific to *A. thaliana*. Given *A. thaliana*'s low heterozygosity²⁸, we assume a haploid host. We begin with an equation for host resistance allele frequency p_t in generation t ,

$$p_t = \frac{p_{t-1}}{p_{t-1} + \theta(1 - p_{t-1})(1 - I)} \quad (1)$$

θ is the ratio of the fitnesses of uninfected susceptible and resistant individuals and I is the fraction of infected susceptible individuals. Infected susceptible individuals are assumed to have zero fitness, although a low nonzero fitness does not affect our results. Importantly, epiphytic populations of *Pseudomonas* on non-host plants contribute to subsequent epidemics²³. Pathogen density z_t during host generation t is determined by

$$z_t = \lambda K(1 - p_{t-1})I + \gamma z_{t-1} \quad (2)$$

Here λ is epiphytic population growth between epidemics, K is host population density, and γ is long-term survival of pathogens. We use an SIR model¹¹ to give the proportion of susceptible individuals that are infected,

$$I = 1 - \exp(R_0 I(1 - p_t) + z_t/N_T) \quad (3)$$

Here R_0 is the fitness of pathogens produced during the epidemic and N_T is the threshold population density.

Re-scaling z_t according to $\hat{z}_t = z_t/N_T$ shows that the dynamics depend only on θ , R_0 and a parameter $\hat{R}_0 = \lambda K/N_T$, the fitness of pathogens from the previous epidemic. Costs of resistance average 7%²²; therefore, we set $\theta = 1/0.93 = 1.08$. We fit the model to the ratio of average pairwise distances within *Rpm1* resistance and susceptibility allelic classes $\pi_R/\pi_S = 5.71$, and to the sampled resistance-allele frequency 0.52. We calculated the expected ratio of average pairwise distances within allelic classes using a standard coalescence formula⁹, into which we substituted harmonic means of resistance and susceptibility allele frequencies from model runs of 10,000 generations (discarding transients). The resulting parameter estimates are $R_0 = 1.35$, $\hat{R}_0 = 2.23$ and $\gamma = 0.2$; the expected ratio of average pairwise distances within allelic classes is 5.7, and the frequency of resistance lies within ± 0.05 of the estimated frequency 13% of the time.

Received 25 March; accepted 7 June 1999.

- Dawkins, R. & Krebs, J. R. Arms race between and within species. *Proc. R. Soc. Lond. B* **205**, 489–511 (1979).
- Kunkel, B. A useful weed put to work: genetic analysis of disease resistance loci in *Arabidopsis thaliana*. *Trends Genet.* **12**, 63–69 (1996).
- Grant, M. R. *et al.* Structure of the *Arabidopsis* *RPM1* gene enabling dual specificity disease resistance. *Science* **269**, 843–846 (1995).
- Jayakar, S. D. A mathematical model for interaction of gene frequencies in a parasite and its host. *Theor. Popul. Biol.* **1**, 140–164 (1970).
- Hamilton, W. D. in *Population Biology of Infectious Diseases* (eds Anderson, R. M. & May, R. M.) 269–296 (Springer, Berlin, 1982).
- Grant, M. R. *et al.* Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **95**, 15843–15848 (1998).
- O'Kane, S. L. & Al-Shehbaz, I. A synopsis of *Arabidopsis* (Brassicaceae). *Novon* **7**, 323–327 (1997).

8. Whalen, M. C., Innes, R. W., Bent, A. F. & Staskawicz, B. J. Identification of *Pseudomonas syringae* pathogens of *Arabidopsis* and a bacterial locus determining avirulence on both *Arabidopsis* and soybean. *Plant Cell* **3**, 49–59 (1991).
9. May, R. M. in *Ecology and Genetics of Host-Parasite Interactions* (eds Rollinson, D. & Anderson, R. M.) 243–262 (Academic, New York, 1985).
10. Maynard Smith, J. & Haigh, J. The hitchhiking effect of a favourable gene. *Genet. Res.* **23**, 23–25 (1974).
11. May, R. M. & Anderson, R. M. Epidemiology and genetics in the coevolution of parasites and hosts. *Proc. R. Soc. Lond. B* **219**, 281–313 (1983).
12. Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites (a review). *Proc. Natl Acad. Sci. USA* **87**, 3566–3573 (1990).
13. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840 (1988).
14. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
15. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
16. McDonald, J. H. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol. Biol. Evol.* **15**, 377–384 (1998).
17. Innan, H., Tajima, F., Terauchi, R. & Miyashita, N. T. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* **143**, 1761–1770 (1996).
18. Kawabe, A., Innan, H., Terauchi, R. & Miyashita, N. T. Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. *Mol. Biol. Evol.* **14**, 1303–1315 (1997).
19. Purugganan, M. D. & Suddith, J. I. Molecular population genetics of the *Arabidopsis* CAULIFLOWER regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. *Proc. Natl Acad. Sci. USA* **95**, 8130–8134 (1998).
20. Purugganan, M. D. & Suddith, J. I. Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. *Genetics* **151**, 839–848 (1999).
21. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* **29**, 208–211 (1989).
22. Nei, M. & Gojibori, T. Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
23. Redei, G. P. *Arabidopsis* as a genetic tool. *Annu. Rev. Genet.* **9**, 111–127 (1986).
24. Bergelson, J. & Purrington, C. Surveying patterns in the cost of resistance in plants. *Am. Nat.* **148**, 536–558 (1996).
25. Hirano, S. S. & Upper, C. D. Population biology and epidemiology of *Pseudomonas syringae*. *Annu. Rev. Phytopath.* **28**, 155–177 (1990).
26. Judson, O. Preserving genes: a model of the maintenance of genetic variation in a metapopulation under frequency-dependent selection. *Genet. Res. Camb.* **65**, 175–191 (1995).
27. Caicedo, A. L., Schaal, B. A. & Kunkel, B. N. Diversity and molecular evolution of the *RPS2* resistance gene in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **96**, 302–306 (1999).
28. Bergelson, J., Stahl, E., Dudek, S. & Kreitman, M. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* **148**, 1311–1323 (1998).
29. Rozas, J. & Rozas, R. DnaSP version 2.0: A novel software package for extensive molecular population genetics analysis. *Comput. Appl. Biosci.* **13**, 307–311 (1997).
30. Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).

Acknowledgements. M. Grant and J. McDowell provided unpublished sequences, and R. Hudson provided computer code. The Nottingham and Ohio State University *Arabidopsis* stock centres provided seeds. We thank R. Hudson, H. Innan, A. Kawabe, C. Langley, Y. Satta and N. Takahata for helpful discussions, and J. Dangel, P. Kareiva, S. Levin and J. McDowell for commenting on earlier versions of the manuscript. G.D. was supported by a Dropkin fellowship and R.M. was supported by an NSF/Sloan Fellowship in Molecular Evolution. This work was supported by an NSF Presidential Award and Packard Fellowship to J.B., an NIH grant to M.K., an NIH grant to J.B. and M.K. and a University of Chicago Hinds Fund grant to E.A.S.

Correspondence and requests for materials should be addressed to J.B. (e-mail: jbergels@midway.uchicago.edu).

Time-dependent reorganization of brain circuitry underlying long-term memory storage

Bruno Bontempi, Catherine Laurent-Demir, Claude Destrade & Robert Jaffard

Laboratoire de Neurosciences Cognitives, CNRS UMR 5807, Université Bordeaux I, Avenue des Facultés, 33405 Talence, France

Retrograde amnesia observed following hippocampal lesions in humans and animals is typically temporally graded^{1,2}, with recent memory being impaired while remote memories remain intact, indicating that the hippocampal formation has a time-limited role in memory storage^{3,4}. However, this claim remains controversial because studies involving hippocampal lesions tell us nothing about the contribution of the hippocampus to memory storage if this region was present at the time of memory retrieval^{5,6}. We therefore used non-invasive functional brain imaging using

(¹⁴C)2-deoxyglucose uptake to examine how the brain circuitry underlying long-term memory storage is reorganized over time in an intact brain. Regional metabolic activity in the brain was mapped in mice tested at different times for retention of a spatial discrimination task. Here we report that increasing the retention interval from 5 days to 25 days resulted in both decreased hippocampal metabolic activity during retention testing and a loss of correlation between hippocampal metabolic activity and memory performance. Concomitantly, a recruitment of certain cortical areas was observed. These results indicate that there is a time-dependent reorganization of the neuronal circuitry underlying long-term memory storage, in which a transitory interaction between the hippocampal formation and the neocortex would mediate the establishment of long-lived cortical memory representations.

The phenomenon of temporally graded retrograde amnesia constitutes a major argument for a memory consolidation process being required for stable memory formation⁷. Although it is generally accepted that this process involves a stabilization of synaptic modifications after a learning experience, regardless of which brain system is involved, it is not yet clear how different systems in an intact brain reorganize over an extended period of time to support long-term memory storage^{8,9}. We therefore mapped (¹⁴C)2-deoxyglucose uptake¹⁰ in selected brain regions of the mouse to measure regional metabolic activity associated with retrieval testing of a recently versus remotely acquired spatial discrimination task. Our objective was to understand the roles of the hippocampal formation and the cortical areas during memory consolidation.

An intrajugular catheter¹¹ was attached to male BALB/c mice that were trained over nine consecutive daily sessions in an eight-arm radial maze to discriminate the three arms that were constantly baited with food from the five arms that were never baited. After a retention interval of either 5 days (5-day retention group; $n = 9$) or 25 days (25-day retention group; $n = 8$), chosen according to data obtained from previous lesion experiments¹², animals were tested once under the same conditions as the initial acquisition. To examine whether changes in functional metabolic activity of the hippocampal formation at the 25-day retention interval are specifically related to the previously learned information, two additional groups of mice were tested under different conditions. Animals were placed either in the same maze as the initial acquisition but with a different location of food (same context/different arms group; $n = 7$) or in a different maze located in a different room (different context/different arms group; $n = 8$) with respect to initial acquisition. Finally, two control groups included mice that were submitted only to the nine initial training sessions (acquisition group; $n = 7$) and mice that were maintained in their home cages ('quiet' control group; $n = 8$) during the experimental period. Levels of regional

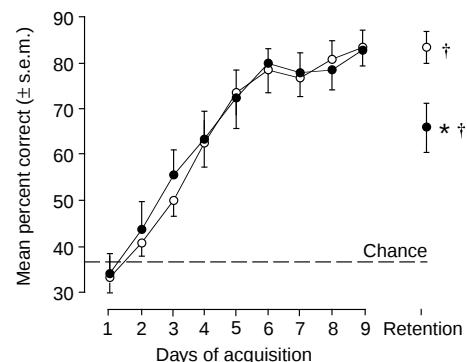


Figure 1 Curves showing the progression of spatial discrimination performance over the nine daily sessions of initial acquisition and the retention performance of the 5-day (open circles) and 25-day (filled circles) retention groups. * $P < 0.05$ as compared to the 5-day retention group; † $P < 0.005$ as compared to chance level.

metabolic activity in 74 brain regions were visualized and measured using (^{14}C)2-deoxyglucose mapping¹⁰.

Discrimination performance of mice in the 5-day retention group was very similar to the level of performance attained on the last acquisition session (Fig. 1). In contrast, performance was worse in subjects in the 25-day retention group (paired Student *t*-test, $t(7) = 2.68$, $P = 0.03$), which, although exhibiting response accuracies significantly above chance level (versus 37.5%; *t*-test, $t(7) = 5.27$, $P < 0.002$), showed a worse performance than the 5-day retention group ($F(1, 16) = 7.67$; $P < 0.02$). A two-way analysis of variance (ANOVA) performed on cerebral metabolic activity data with the 5-day and 25-day retention intervals as between-group factor and the 74 selected brain regions as within-subjects factor indicated a highly significant group $\times$ region interaction ($F(73, 1095) = 14.61$; $P < 0.0001$), showing that region-specific changes in metabolic activity occurred between these two retention intervals. The same analysis revealed no such region-specific changes in regional metabolic activity between the acquisition group trained for 9 days and the 5-day retention group (group $\times$ region interaction, $F(73, 1022) = 1.23$; $P > 0.10$). Analysis of individual regions of the hippocampal formation revealed a significantly lower level of metabolic activity in the 25-day retention group with respect to both the acquisition and 5-day retention groups (Figs 2 and 3). Changing either the reward location (same context/different arms group) or the context (different context/different arms group) at the 25-day retention interval was associated with a re-activation of functional metabolic activity in the hippocampal formation, indicating that the decrease in hippocampal metabolic activation at the longer retention interval may have been related to the retrieval of information acquired 25 days earlier. Conversely, this analysis also revealed that lengthening the retention interval from 5 days to 25 days resulted in a significant increase in the metabolic activity of certain cortical structures. The largest increases were observed in the frontal, anterior cingulate and

temporal cortices (Figs 2 and 3), indicating greater recruitment of these regions in the retrieval of remote as compared to recently acquired information. The pattern of changes observed at the 25-day retention interval in the cingulate cortex was opposite to that of the 5-day interval (Fig. 3), in which the anterior division exhibited a significant increase in metabolic activity, whereas the posterior division exhibited a significant decrease. These differential changes are in agreement with previous findings indicating a dissociation between these two cortices, with the posterior division being functionally and neuroanatomically related to the hippocampal formation and the anterior division being related to the frontal region^{13,14}.

The decrease of metabolic activation in the hippocampal formation at the 25-day interval as compared to the 5-day interval indicates that discrimination performance at the 25-day interval is less dependent on the functional activation of the hippocampal formation. The level of metabolic activity observed in the hippocampal formation at the 25-day interval was significantly less than that in the hippocampal formation of 'quiet' control animals. This finding indicates that the reduction in hippocampal metabolic activation is not solely the consequence of a reduction in the size of the neuronal population representing the familiar (consolidated) information, or of more widely distributed hippocampal networks subserving information storage⁵, but also reflects the existence of inhibitory influences from brain regions that are actively engaged in memory retrieval¹⁵.

A factor analysis carried out at each retention interval, including both response accuracy and levels of metabolic activity in selected brain regions, confirms this finding and further supports the idea that the hippocampal formation has a transitory role in memory storage. Response accuracy at the 5-day retention interval appeared in the same cluster as the hippocampal formation (Fig. 4a), whereas cortical areas were relatively dispersed and appeared outside the hippocampal cluster. The same analysis performed in mice tested at

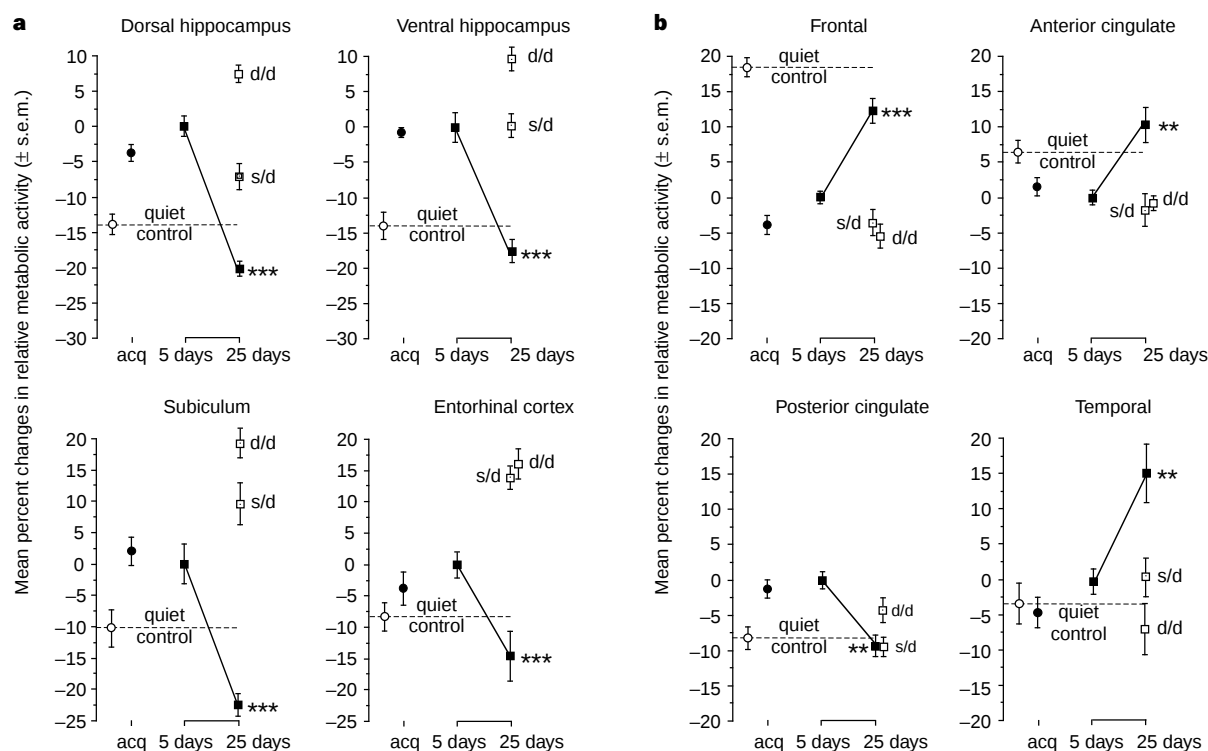


Figure 2 Regional metabolic activity measured in the brains of mice as a function of the retention interval (5 days or 25 days) following initial acquisition (acq) of the spatial discrimination task. **a**, Hippocampal formation; **b**, cortical areas. Testing was conducted at the 25-day interval either with the same conditions as initial acquisition, or with a change in reward contingencies: same context/different

arms (s/d) or different context/different arms (d/d). Results are expressed as mean percentage changes in relative metabolic activity as compared to that observed at the 5-day retention interval. $**P < 0.01$ and $***P < 0.001$ as compared to the 5-day interval.

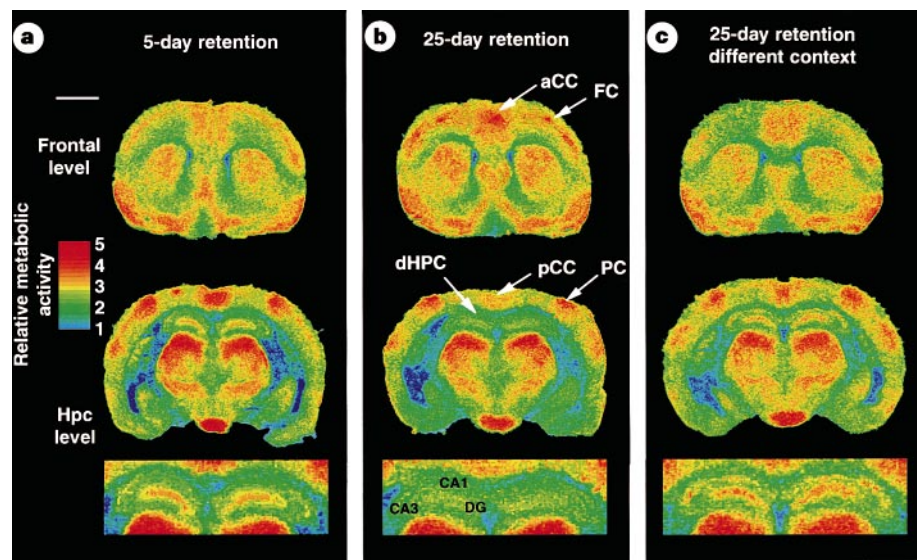


Figure 3 (^{14}C)2-deoxyglucose colour-coded autoradiographs. These were obtained from frontal brain sections of mice that had initially undergone 9 days of spatial discrimination testing and were then submitted to retention testing at either 5 days (**a**) or 25 days (**b**) after initial acquisition; **c**, mice were submitted 25 days after acquisition to a different set of baited arms in a different context. Top

and middle, sections taken at the frontal and dorsal hippocampal (dHpc) levels, respectively. Bottom, a magnified view of the dorsal hippocampus including the CA1, CA3 and dentate gyrus (DG). aCC, anterior cingulate cortex; FC, frontal cortex; pCC, posterior cingulate cortex; PC, parietal cortex. Scale bar, 2 mm.

the 25-day interval (Fig. 4b) provided a different picture. Specifically, response accuracy no longer appeared in the hippocampal formation cluster but was instead part of a cluster that included cortical areas extending from the frontal to the anterior cingulate cortices. The correlation matrix table for each retention interval is summarized in Fig. 4c.

Taken together, our results indicate a time-dependent reorganization of the neuronal circuitry underlying long-term memory storage. As time passes after learning, we observe a shift in the relative roles played by different cerebral structures engaged in the processing and retrieval of information, a concept initially inferred from lesion studies^{1–4}. Memory performance is strongly linked to metabolic activation of the hippocampal formation shortly after training. However, as memory consolidation proceeds, its functional contribution diminishes with regions outside the hippocam-

pal formation, namely cortical areas including the frontal, anterior cingulate and temporal cortices, becoming capable of mediating the retrieval of previously learned information. These findings do not support the recent idea of multiple memory traces being formed and distributed within the hippocampal formation, and remaining a necessary component of a memory representation as long as they are viable⁵. In contrast, our findings provide further evidence in support of the recently questioned model of memory consolidation that postulates a gradual reorganization of the neural substrates underlying long-term memory storage^{6,9,16,17}. According to this model, the consolidation process would be accomplished by means of a transitory interaction between the hippocampal formation and the neocortex to establish permanent cortical memory representations⁹.

Changing reward contingencies at the 25-day retention interval

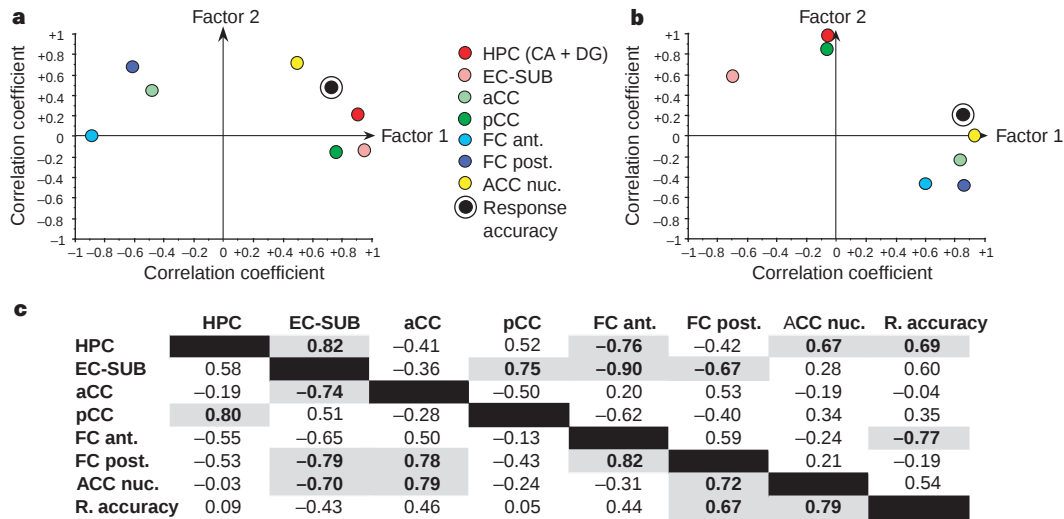


Figure 4 Factor analysis showing clustering of response accuracy and metabolic activity in mouse brain regions during retention testing. **a**, Retrieval performance of recent information (5 days retention) appears in the hippocampal cluster. **b**, Retrieval performance of remote memories (25 days retention) now appears in the cortical cluster. **c**, Correlation matrix table. Correlation coefficients for the 5- and

25-day retention groups are presented above and below the solid black line, respectively. Shaded boxes show significant correlations ($P < 0.05$). HPC, hippocampus; EC-SUB, entorhinal cortex/subiculum; aCC and pCC, anterior and posterior cingulate cortices; FC ant. and FC post., anterior and posterior parts of the frontal cortex; ACC nuc., accumbens nucleus.

resulted in the reactivation of the hippocampal formation. This finding suggests that the hippocampal formation is required to encode new information¹⁸ or to maintain recently acquired information in memory until the consolidation process has ended. Our data also show that the retrieval of remotely acquired information involves the participation of certain neocortical areas. The increased metabolic activation in the frontal cortex, together with the recruitment of the anterior cingulate and temporal cortices observed at the 25-day interval, may reflect the participation of these areas in the organized triggering, access to and use of previously stored spatial representations^{19–21}. In humans, participation of the cingulate and temporal cortices has been described during the retrieval of autobiographic memory²². Prefrontal activation has also been observed in brain imaging studies of explicit retrieval, and has been related to the temporary sequencing and organization of information and to the effort of attempting to recall previously studied items^{23–26}. Participation in such effortful recall could account for the greater frontal metabolic activity observed at the 25-day interval relative to the 5-day interval, as mice tested at the longer delay may have to deploy more retrieval effort to access previously learned information than do mice tested shortly after training. The higher level of recruitment of neocortical areas during retrieval of remote as compared to recent memories points to the existence of dynamic interactions between the hippocampal formation and cortical areas during the course of the consolidation process^{6,9,11,27}. □

Methods

Animals. Male mice (12–14 weeks old) of the BALB/c By JICO inbred strain were used. Two weeks before behavioural testing, a polyethylene catheter was inserted and secured in the jugular vein of the mice to allow intravenous injection of the (¹⁴C)-2-deoxyglucose tracer^{11,14}. All surgical and experimental procedures were conducted in accordance with French regulations for the care and use of laboratory animals.

Behavioural testing procedure. The automated eight-arm radial maze used for spatial discrimination testing¹¹ was made of grey plexiglas and consisted of a circular central platform (30 cm diameter) from which radiated in a symmetrical fashion eight arms (50 cm long and 11 cm wide). A food-pellet tray was situated at the end of each arm. Animals were food deprived to maintain their body weight at 90% of their *ad libitum* weight during testing, and were initially allowed free exploration sessions on two successive days with all arms baited with food to familiarize them with the maze and its environment. For spatial discrimination testing, each subject was assigned a different set of three constantly baited arms (arms 1, 4, 6) with the specification that the sequence of angles between the baited arms was 135°–90°–135°. Mice were tested for acquisition of this reference memory task over nine daily sessions composed of six trials separated by an interval of 1 min. Each trial started by the opening of all eight doors simultaneously while the animal was on the central platform. After each visit to an arm, the door giving access to that arm was automatically closed, preventing the subject from making working memory errors (re-entries into an already visited arm). Each trial was terminated after the animal retrieved the third food pellet and had returned to the central platform. Discrimination performance was assessed by the percentage of correct responses recorded for the first three arm choices (3 choices per trial × 6 trials per session = 18 choices). Similar rates of acquisition and similar levels of performance attained on the ninth session were observed for the five trained groups (*n* = 39). The maze was located in a soundproofed room that was decorated with pictures and objects to allow spatial orientation¹¹. A second maze, identical to the first but located in a different room, was used to evaluate the effect of a change of context on brain metabolic activity at the 25-day retention interval.

(¹⁴C)-2-deoxyglucose procedure. (¹⁴C)-2-deoxyglucose was injected 1 min before the beginning of either the last training session or the retention sessions (5 days and 25 days after training). Mice were killed at the end of each 35-min session and their brains were rapidly removed, frozen in liquid nitrogen and prepared for autoradiography¹¹. Autoradiographs were analysed by semi-quantitative densitometry using a BIOCROM 200 computerized image-processing system. We selected for analysis 74 brain regions anatomically defined

according to the atlas of Lehman²⁸. The mean optical density of each structure was measured bilaterally using a minimum of four consecutive sections in each brain by two different experimenters who were blind to the experimental conditions. The value for each structure was expressed as the ratio of the optical density of that structure and the average optical density for all regions measured for each animal (whole-brain ratio)²⁹. Thus, this optical-density ratio was taken as a measure of the relative metabolic activity of each structure. The mean (¹⁴C)-2-deoxyglucose uptake ratios contributed by each animal in a group were averaged to generate the final means. Statistical analyses were performed using ANOVA.

Factor analysis. This analysis was used to determine the manner in which specific brain regions aggregated (clusters of regions that displayed highly correlated metabolic activity) and the manner in which metabolic activity in these brain regions correlated with response accuracy as a function of the retention interval (5 or 25 days). The inter-relations among the variables were taken into consideration and several independent factors were formed to account for the variance in the data. Thus, if response accuracy at a given retention interval was related to functional activity of certain brain regions, then both of these measures would have high loadings on the same factor. For the factor analysis, the metabolic activity and response accuracy data obtained at the 5-day and 25-day retention intervals were pooled into a correlation matrix using the Statview 4.5 software (Abacus Concepts) and submitted to a Solution–Varimax orthogonal transformation. In addition to the hippocampal formation and neocortical regions, the nucleus accumbens was included in the analysis.

Received 7 April; accepted 9 June 1999.

1. Reed, J. M. & Squire, L. R. Retrograde amnesia for facts and events: findings from four new cases. *J. Neurosci.* **18**, 3943–3954 (1998).
2. Squire, L. R. Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychol. Rev.* **99**, 195–231 (1992).
3. Zola-Morgan, S. M. & Squire, L. R. The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science* **250**, 288–290 (1990).
4. Anagnostaras, S. G., Maren, S. & Fanselow, M. S. Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: within-subjects examination. *J. Neurosci.* **19**, 1106–1114 (1999).
5. Moscovitch, M. & Nadel, L. Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Curr. Opin. Neurobiol.* **8**, 297–300 (1998).
6. Knowlton, B. J. & Fanselow, M. S. The hippocampus, consolidation and on-line memory. *Curr. Opin. Neurobiol.* **8**, 293–296 (1998).
7. Polster, M. R., Nadel, L. & Schacter, D. L. Cognitive neuroscience analysis of memory: a historical perspective. *J. Cogn. Neurosci.* **3**, 95–116 (1991).
8. Dudai, Y. Consolidation: fragility on the road to the engram. *Neuron* **17**, 367–370 (1996).
9. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
10. Sokoloff, L. et al. The (¹⁴C)-deoxyglucose method for measurement of local cerebral glucose utilization: theory, procedure and normal values in the conscious and anesthetized albino rat. *J. Neurochem.* **28**, 897–916 (1977).
11. Bontempi, B., Jaffard, R. & Destrade, C. Differential temporal evolution of post-training changes in regional brain glucose metabolism induced by repeated spatial discrimination training in mice: visualization of the memory consolidation process? *Eur. J. Neurosci.* **8**, 2348–2360 (1996).
12. Cho, Y. H., Beracoche, D. & Jaffard, R. Extended temporal gradient for the retrograde and anterograde amnesia produced by ibotenate entorhinal cortex lesions in mice. *J. Neurosci.* **13**, 1759–1766 (1993).
13. Finch, D. M., Derian, E. L. & Babb, T. L. Afferent fibers to rat cingulate cortex. *Exp. Neurol.* **83**, 468–485 (1984).
14. Sif, J., Meunier, M., Messier, C., Calas, A. & Destrade, C. Quantitative (¹⁴C)-2-deoxyglucose study of a functional dissociation between anterior and posterior cingulate cortices in mice. *Neurosci. Lett.* **101**, 223–228 (1989).
15. Nyberg, L. et al. Network analysis of positron emission tomography regional cerebral blood flow data: ensemble inhibition during episodic memory retrieval. *J. Neurosci.* **16**, 3753–3759 (1996).
16. Teyler, T. J. & DiScenna, P. The hippocampal memory indexing theory. *Behav. Neurosci.* **100**, 147–154 (1986).
17. Damasio, A. R. Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition* **33**, 25–62 (1989).
18. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
19. Meunier, M., Jaffard, R. & Destrade, C. Differential involvement of anterior and posterior cingulate cortices in spatial discrimination learning in a T-maze in mice. *Behav. Brain Res.* **44**, 133–143 (1991).
20. Markowitsch, H. J. Which brain regions are critically involved in the retrieval of old episodic memory. *Brain Res. Rev.* **21**, 117–127 (1995).
21. Buckner, R. L., Kelley, W. M. & Petersen, S. E. Frontal cortex contributes to human memory formation. *Nature Neurosci.* **2**, 311–314 (1999).
22. Fink, G. R. et al. Cerebral representation of one's own past: neural networks involved in autobiographical memory. *J. Neurosci.* **16**, 4275–4282 (1996).
23. Ungerleider, L. G. Functional brain imaging studies of cortical mechanisms for memory. *Science* **270**, 769–775 (1995).
24. Schacter, D. L., Alpert, N. M., Savage, C. R., Rauch, S. L. & Albert, M. S. Conscious recollection and the human hippocampal formation: evidence from positron emission tomography. *Proc. Natl Acad. Sci. USA* **93**, 321–325 (1996).
25. Gabrieli, J. D. E. Cognitive neuroscience of human memory. *Annu. Rev. Psychol.* **49**, 87–115 (1998).
26. Alkire, M. T., Jaffard, R. J., Fallon, J. H. & Cahill, L. Hippocampal, but not amygdala, activity at encoding correlates with long-term, free recall of nonemotional information. *Proc. Natl Acad. Sci. USA* **95**, 14506–14510 (1998).

27. Buzsaki, G. Memory consolidation during sleep: a neuropsychological perspective. *J. Sleep Res.* 7, 17–23 (1998).
28. Lehman, A. *Atlas Stéréotaxique du Cerveau de la Souris* (Editions du CNRS, Paris, 1974).
29. McIntosh, A. R. & Gonzalez-Lima, F. Network interactions among limbic cortices, basal forebrain and cerebellum differentiate a tone conditioned as a Pavlovian excitator or inhibitor: fluorodeoxyglucose mapping and covariance structural modelling. *J. Neurophysiol.* 72, 1717–1733 (1994).

Acknowledgements. We thank T. Durkin, B. Poucet and Y. H. Cho for comments on earlier drafts of the manuscript. This work was funded by grants from the CNRS and the Université Bordeaux I.

Correspondence and requests for materials should be addressed to B.B. (e-mail: b.bontempi@neurocog.u-bordeaux.fr).

Memory for places learned long ago is intact after hippocampal damage

Edmond Teng* & Larry R. Squire*†‡§

† Veterans Affairs Medical Center, San Diego, California 92161, USA
Departments of ‡ Psychiatry, * Neurosciences and § Psychology,
University of California, San Diego, La Jolla, California 92093, USA

The hippocampus is part of a system of structures in the medial temporal lobe that are essential for memory^{1–3}. One influential view of hippocampal function emphasizes its role in the acquisition and retrieval of spatial knowledge^{4,5}. By this view, the hippocampus constructs and stores spatial maps and is therefore essential for learning and remembering places, including those learned about long ago. We tested a profoundly amnesic patient (E.P.), who has virtually complete bilateral damage to the hippocampus and extensive damage to adjacent structures in the medial temporal lobe. We asked him to recall the spatial layout of the region where he grew up, from which he moved away more than 50 years ago. E.P. performed as well as or better than age-matched control subjects who grew up in the same region and also moved away. In contrast, E.P. has no knowledge of his current neighbourhood, to which he moved after he became amnesic. Our results show that the medial temporal lobe is not the permanent repository of

spatial maps, and support the view that the hippocampus and other structures in the medial temporal lobe are essential for the formation of long-term declarative memories, both spatial and non-spatial, but not for the retrieval of very remote memories, either spatial or non-spatial^{3,6}.

Patient E.P., a 76-year-old former laboratory technician, became amnesic in 1992 after an episode of herpes simplex encephalitis. He has extensive bilateral damage to all the components of the memory system of the medial temporal lobe, including the hippocampus⁷ (Fig. 1). E.P.'s amnesia is so severe that he fails to recognize his examiners, even after 40 visits to his house in one year, and he scores at chance on a variety of verbal and non-verbal tasks of recognition memory^{8,9}. He also has severe difficulty recalling facts and events that occurred during the 40 years before his illness⁷.

E.P. grew up in the Hayward-Castro Valley area of California during the 1930s and 1940s (Fig. 2). He moved away from the area at the age of 28 and has returned only occasionally. We identified five other individuals who attended E.P.'s high school during this same period, lived there for about as long as E.P. did (controls, 26 years; E.P., 22 years), and then moved away from the area. We tested all six subjects on four tests of topographical memory that assessed their spatial knowledge of the region in which they grew up. They were asked to describe how they would navigate from their homes to different locations in the area (familiar navigation), between different locations in the area (novel navigation), and between these same locations if a main street were blocked off (alternative routes). They were also asked to imagine themselves in a particular orientation at certain locations and then to point towards specific landmarks (pointing to landmarks).

E.P. scored well on all four tests (Fig. 3). His poorer performance on the first administration of 'pointing to landmarks' was due to

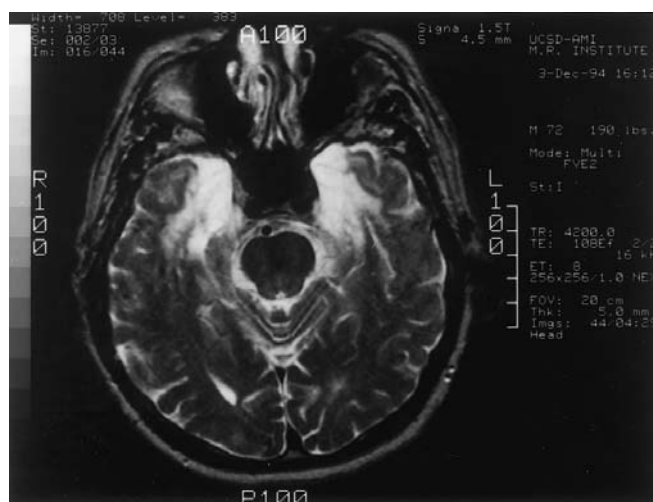


Figure 1 An axial T2 weighted image through the temporal lobes of patient E.P. The extent of damage to the bilateral medial temporal lobe can be seen; it extends caudally from the temporal pole and includes the hippocampal region (dentate gyrus, cell fields of the hippocampus proper and subicular complex), the entorhinal cortex, the perirhinal cortex, the parahippocampal cortex and the amygdaloid complex. The lesion also extends laterally to include the anterior portion of the fusiform gyrus. Damage to the hippocampal region is virtually complete, except for a small fragment of tissue in the lateral ventricle bilaterally. The adjacent cortical structures of the medial temporal lobe are extensively damaged, possibly sparing the caudal portion of the parahippocampal cortex.



Figure 2 Street map of a portion of the Hayward-Castro Valley region from the 1940s (Thomas Bros. Map Co.). The locations of four representative landmarks used in the topographical memory tasks are shown: A, Bret Harte School; B, Hayward Union High School; C, Hayward Theatre; D, Castro Valley Grammar School. The locations used in the four topographical memory tasks encompassed an area of approximately 50 square miles, a larger area than is shown here.

two instances in which he reported the corrected heading verbally (for example, southwest) but then pointed in a different direction. When these two questions were omitted, his median error on this first administration of the task was 35 degrees. Based on all the questions in both testing sessions, his median pointing error was 38 degrees (control median, 34 degrees), and his mean heading vector relative to the correct heading was significantly above chance ($P < 0.05$, Rayleigh's test¹⁰) and not significantly different from the correct heading ($P > 0.10$, Stephen's test¹⁰). The mean heading vectors for four of the five control subjects were also significantly above chance and were statistically indistinguishable from the correct heading.

In Fig. 3a, responses in the verbal navigation tasks were scored as correct if they included the correct sequence and direction of turns, even if specific street names were not recalled. E.P.'s performance was indistinguishable from that of the controls. When a different scoring criterion was used, which required both the correct sequence of turns and the correct street names, E.P. continued to perform either above or within the control range on all three verbal navigation tasks (E.P. scored 57%, 75% and 38% against control ranges of 50–100%, 20–71% and 25–100%, respectively).

Although the specific questions used to assess topographical memory differed across subjects (because not all subjects were familiar with exactly the same locations), there were 3–6 questions for each test that were given to E.P. and at least four of the five control subjects. For this common set of questions, E.P. averaged 70% correct on the 'familiar navigation' task (control mean, 80%), 92% correct on the 'novel navigation' task (control mean, 81%) and 88% correct on the 'alternative routes' task (control mean, 88%). On the 'pointing to landmarks' task, E.P.'s average pointing error was 40 degrees (control mean, 36 degrees).

According to the spatial hypothesis of hippocampal function, the spatial maps stored in the hippocampus enable flexible navigation by encoding several routes to the same destination⁴. However, E.P.'s performance on the 'alternative routes' task was slightly better than that of the control subjects, indicating that flexible navigation through environments learned long ago is possible despite virtually complete hippocampal damage.

We also asked all six subjects to describe how they would navigate from the homes in which they currently reside to each of five different locations in the neighbourhood. All control subjects were able to provide accurate directions in response to each question (100% correct), but E.P., who moved to San Diego County in 1993, after he became amnesic, was unable to provide any response at all

to any of the questions (0% correct).

The idea that the hippocampus or related structures in the medial temporal lobe are important for spatial cognition is based on the discovery of place cells in the rodent hippocampus¹¹, findings of impaired spatial learning in human patients^{12–14} and animals^{5,15} with hippocampal damage, and neuroimaging studies that find activation of these regions in tasks of spatial navigation^{16–19}. E.P.'s failure to acquire new spatial knowledge about his current neighbourhood is consistent with these findings. It has also been supposed that damage to the medial temporal lobe might impair all autobiographical memories and spatial memories, regardless of when they were acquired²⁰. Yet patient E.P., despite almost complete damage to his medial temporal lobe, has intact memory for remote autobiographical episodes⁷ and, as shown here, has intact memory for remote spatial information. Our findings therefore support the view that the medial temporal lobe is essential for the formation of long-term declarative memories, both spatial and non-spatial, but not for the retrieval of very remote spatial or non-spatial memories^{3,6}. Case studies of patients with topographical amnesia, defined²¹ as "when the patient loses his bearings in a well-known environment and is not able to give a verbal or graphic description of familiar routes or places", have consistently reported damage to posterior cortical regions, particularly the parietal and retrosplenial cortices^{22–27}, indicating that these may be the brain regions that maintain long-term spatial maps of learned environments. □

Methods

E.P.'s performance was compared with the performance of healthy subjects (three men and two women) who attended the Hayward Union High School (Hayward, California) at the same time as E.P. They were matched to E.P. with respect to age (control mean, 75 years; E.P., 76 years) and duration of education (12 years). They had lived in the Hayward-Castro Valley region for an average of 26 years (range, 11–52 years; E.P., 22 years), and had then lived elsewhere in California for an average of 44 years (range 24–50 years; E.P., 51 years). They have occasionally returned to the area for visits but, with the exception of one control subject, not within the past two years. E.P.'s most recent visit to the region was in 1993. The five controls had intact memory functions (mean Wechsler Memory Scale-III subtest scores were 11.0 for Logical Memory and 10.6 for Verbal Paired Associates with an age-matched score for the normative sample of 10.0) (ref. 28).

E.P. was tested twice, once before any of the control subjects were tested and again after all the control subjects were tested. Through interviews before formal testing, approximately 20 familiar locations in the Hayward-Castro Valley region of California in the 1930s and 1940s were identified for each subject. These locations were then used to construct four tasks.

In the 'familiar navigation' task, subjects were asked to describe a route from their childhood home to different locations in the region (E.P. was asked 10 questions; control mean, 9.8); for example: "Can you tell me how to get from your house to the Bret Harte School?". In the 'novel navigation' task, subjects were asked to describe a route from one location to another (E.P., 8 questions; control mean, 7.8); for example: "Can you tell me how to get from the Hayward Union High School to the Hayward Theatre?". Next, after navigating a route between these two locations, subjects were asked to find an alternative route to the same destination, assuming that one of the major streets that they had just used was now blocked off (alternative routes task). Responses to these questions were scored using printed transcripts of the testing sessions. For example, in patient E.P.'s transcript, his response to the 'novel navigation' question was: "Make a left out of the high school. To A Street. Make a right on A Street, down to Castro. A left on Castro, and then in the middle of the block is the Hayward Theatre." Finally, in the 'pointing to landmarks' task, subjects were asked to imagine themselves in a given location, facing a particular direction, and then to point in the direction of another location (E.P., 9 questions; control mean, 8.6); for example: "You are standing in front of the Hayward Union High School, facing Foothill Boulevard. Can you point in the direction of the Castro Valley Grammar School?" Performance on this task was measured as the error (in degrees) between the correct heading and the direction to which the subject pointed.

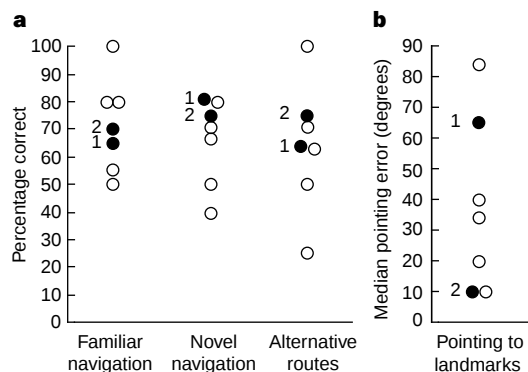


Figure 3 Performance on four tasks of topographical memory. Open circles, scores of five control subjects; filled circles, E.P.'s scores on two different occasions separated by 9 months. Numbers next to the filled circles indicate scores for E.P.'s first and second test sessions. **a**, Percentage correct on three verbal navigation tasks that required negotiating either familiar routes, novel routes or alternative routes (when the most direct route was blocked). **b**, Median error in degrees on a task in which subjects pointed to particular locations while imagining themselves oriented at other locations in the area.

Although the specific locations used in these tests varied between subjects, 3–6 of the same questions on each test were administered to E.P. and at least four of the five controls. In addition, the 'familiar navigation' and 'novel navigation' tasks were matched across subjects with respect to the distance travelled (mean, 3.0 and 3.3 miles, respectively) and the number of turns needed (mean, 3.0 and 2.7 turns, respectively) to reach each destination. The routes that subjects reported in the verbal navigation tasks were scored as correct if they incorporated the correct sequence and direction of turns necessary to reach the destination. All subjects typically reported street names as they navigated their routes. However, presumably because of his anomia (Boston Naming Test score, 42; maximum score, 60, mean of four control subjects, 54.5)⁷, E.P. omitted street names more frequently than did the control subjects (0.5 omissions per question compared with 0.3 omissions for controls; range, 0.1–0.5). Accordingly, we also used another scoring method, which required both a correct sequence of turns and correct street names. An independent scorer who was blind to subject identity scored all transcripts using both scoring criteria. Average inter-scorer reliability across both scoring criteria was 0.91.

The five verbal navigation questions for current neighbourhoods were administered in the same way as the 'familiar navigation' task. Across subjects, the questions were similar with respect to the distance travelled and the number of turns needed to reach location (mean, 6.5 miles and 5.5 turns, respectively).

Received 29 April; accepted 9 June 1999.

- Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatr.* **20**, 11–21 (1957).
- Mishkin, M. Memory in monkeys severely impaired by combined but not separate removal of the amygdala and hippocampus. *Nature* **273**, 297–298 (1978).
- Squire, L. R. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231 (1992).
- O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, Oxford, 1978).
- Jarrard, L. E. What does the hippocampus really do? *Behav. Brain Res.* **71**, 1–10 (1995).
- Cohen, N. J. & Eichenbaum, H. *Memory, Amnesia, and the Hippocampal System* (MIT Press, Cambridge, Massachusetts, 1993).
- Reed, J. M. & Squire, L. R. Retrograde amnesia for facts and events: Findings from four new cases. *J. Neurosci.* **18**, 3943–3954 (1998).
- Hamann, S. B. & Squire, L. R. Intact perceptual memory in the absence of conscious memory. *Behav. Neurosci.* **111**, 850–854 (1997).
- Reed, J. M., Hamann, S. B., Stefanacci, L. & Squire, L. R. When amnesic patients perform well on recognition memory tasks. *Behav. Neurosci.* **111**, 1163–1170 (1997).
- Batschelet, E. *Statistical Methods for the Analysis of Problems in Animal Orientation and Certain Biological Rhythms* (American Institute of Biological Sciences, Washington DC, 1965).
- O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
- Smith, M. L. & Milner, B. The role of the right hippocampus in the recall of spatial location. *Neuropsychologia* **19**, 781–793 (1981).
- Cave, C. B. & Squire, L. R. Equivalent impairment of spatial and nonspatial memory following damage to the human hippocampus. *Hippocampus* **1**, 329–340 (1991).
- Maguire, E. A., Burke, T., Phillips, J. & Staunton, H. Topographical disorientation following unilateral temporal lobe lesions in humans. *Neuropsychologia* **34**, 993–1001 (1996).
- Morris, R. G. M., Garrud, P., Rawlins, J. N. P. & O'Keefe, J. Place navigation in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
- Aguirre, G. K., Detre, J. A., Alsop, D. C. & D'Esposito, M. D. The parahippocampus subserves topographical learning in man. *Cereb. Cort.* **6**, 823–829 (1996).
- Maguire, E. A., Frackowiak, R. S. J. & Frith, C. D. Learning to find your way: A role for the human hippocampal formation. *Proc. R. Soc. Lond.* **263**, 1745–1750 (1996).
- Maguire, E. A., Frackowiak, R. S. J. & Frith, C. D. Recalling routes around London: Activation of the right hippocampus in taxi drivers. *J. Neurosci.* **17**, 7103–7110 (1997).
- Maguire, E. A. *et al.* Knowing where and getting there: A human navigation network. *Science* **280**, 921–924 (1998).
- Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
- De Renzi, E., Faglioni, P. & Villa, P. Topographical amnesia. *J. Neurol. Neurosurg. Psychiatr.* **40**, 498–505 (1977).
- Pateron, A. & Zangwill, O. L. A case of topographical disorientation associated with a unilateral cerebral lesion. *Brain* **68**, 188–211 (1945).
- Levine, D. N., Warach, J. & Farah, M. Two visual systems in imagery: Dissociation of "what" and "where" in imagery disorders due to bilateral posterior cerebral lesions. *Neurology* **35**, 1010–1018 (1985).
- Bottini, G., Cappa, S., Geminiani, G. & Sterzi, R. Topographic disorientation—a case report. *Neuropsychologia* **28**, 309–312 (1990).
- Incisa della Rochetta, A., Cipolletti, L. & Warrington, E. K. Topographical disorientation: Selective impairment of locomotor space? *Cortex* **32**, 727–735 (1996).
- Pai, M.-C. Topographic disorientation: Two cases. *J. Formos. Med. Assoc.* **96**, 660–663 (1997).
- Takahashi, N., Kawamura, M., Shiota, J., Kasahata, N. & Hirayama, K. Pure topographic disorientation due to right retrosplenial lesion. *Neurology* **49**, 464–469 (1997).
- Wechsler, D. *Wechsler Memory Scale-III. Administration and Scoring Manual* (Psychological Corporation, San Antonio, Texas, 1997).

Acknowledgements. We thank J. Zouounis, L. Stefanacci, J. Frascino and the Hayward Area Historical Society for assistance. This work was supported by the Medical Research Service of the Department of Veterans Affairs, NIMH, and the McDonnell-Pew Center for Cognitive Neuroscience.

Correspondence and requests for materials should be addressed to L.R.S. (e-mail: lsquire@ucsd.edu).

A basal ganglia pacemaker formed by the subthalamic nucleus and external globus pallidus

Dietmar Plenz* & Stephen T. Kitai

Department of Anatomy and Neurobiology, University of Tennessee, College of Medicine, Memphis, Tennessee 38163, USA

The subthalamic nucleus of the basal ganglia (STN) is important for normal movement^{1,2} as well as in movement disorders^{3–5}. Lesioning⁶ or deep-brain stimulation^{7,8} of the STN can alleviate resting tremor in Parkinson's disease. The STN⁵ and its target nuclei^{9,10} display synchronized oscillatory burst discharge at low frequencies, some of which correlate with tremor, but the mechanism underlying this synchronized bursting is unknown. Here we show that the excitatory STN and inhibitory, external globus pallidus (GPe) form a feedback system that engages in synchronized bursting. In mature organotypic cortex–striatum–STN–GPe cultures, neurons in the STN and GPe spontaneously produce synchronized oscillating bursts at 0.4, 0.8 and 1.8 Hz. Pallidal lesion abolishes this bursting, whereas cortical lesion favours bursting at 0.8 Hz. Pallidal bursts, although weaker than STN bursts, were required for synchronized oscillatory burst generation by recruitment of subthalamic rebound excitation. We propose that the STN and GPe constitute a central pacemaker modulated by striatal inhibition of GPe neurons. This pacemaker could be responsible for synchronized oscillatory activity in the normal and pathological basal ganglia.

To test our proposal that the STN and GPe produce synchronized oscillatory bursts, we developed an *in vitro* model¹¹ in which both nuclei were co-cultured with the cortex and striatum, their main extrinsic input sources, to ensure proper maturation. The STN and GPe were obtained from rats at postnatal day 0–2 and cultured with frontomedial cortex and dorsolateral striatum. After 38 ± 1 days *in vitro* ($n = 58$ cultures), spontaneous single- and multi-unit activities were recorded from the STN and GPe with one or two extracellular electrodes.

Spontaneous activity in the STN showed distinctive, stereotypic periods of oscillatory burst discharge that lasted for 10–15 s (Fig. 1a). Intraburst firing rates reached several hundred spikes per second, and the bursts oscillated at low frequencies. Between bursts, STN units were either silent or fired irregularly at low rates. The burst activity of STN units was phase-locked and synchronized with other STN and GPe units (Fig. 1b, c), showing that it reflects population activity across both nuclei. Spontaneous synchronized bursting occurred regularly every 1–2 min and with occasional shifts in main frequency (Fig. 1d).

Based on correlation analysis and frequency plots using continuous periods of spontaneous spiking (324 ± 135 s per neuron), about half of STN (83/181) and a third of GPe units (31/102) fired in oscillatory bursts with frequencies between 0.1 and 4 Hz (20 ms time resolution). Similarly, 61% of STN–STN (46/76), 44% of STN–GPe (33/75), and 23% of GPe–GPe (4/17) neuronal pairs displayed synchronized oscillatory bursts in that frequency range.

The STN–GPe system showed clear preferences for particular frequencies during synchronized bursts. The relative power spectrum analysis revealed two main population frequencies at $f_{01} = 0.44$ Hz and $f_{02} = 0.79$ Hz, respectively (Fig. 2a). A third

* Present address: Unit of Neural Network Physiology, Laboratory of Systems Neuroscience, National Institute of Mental Health, Bldg 36 2D-30, Bethesda, Maryland 20892, USA.

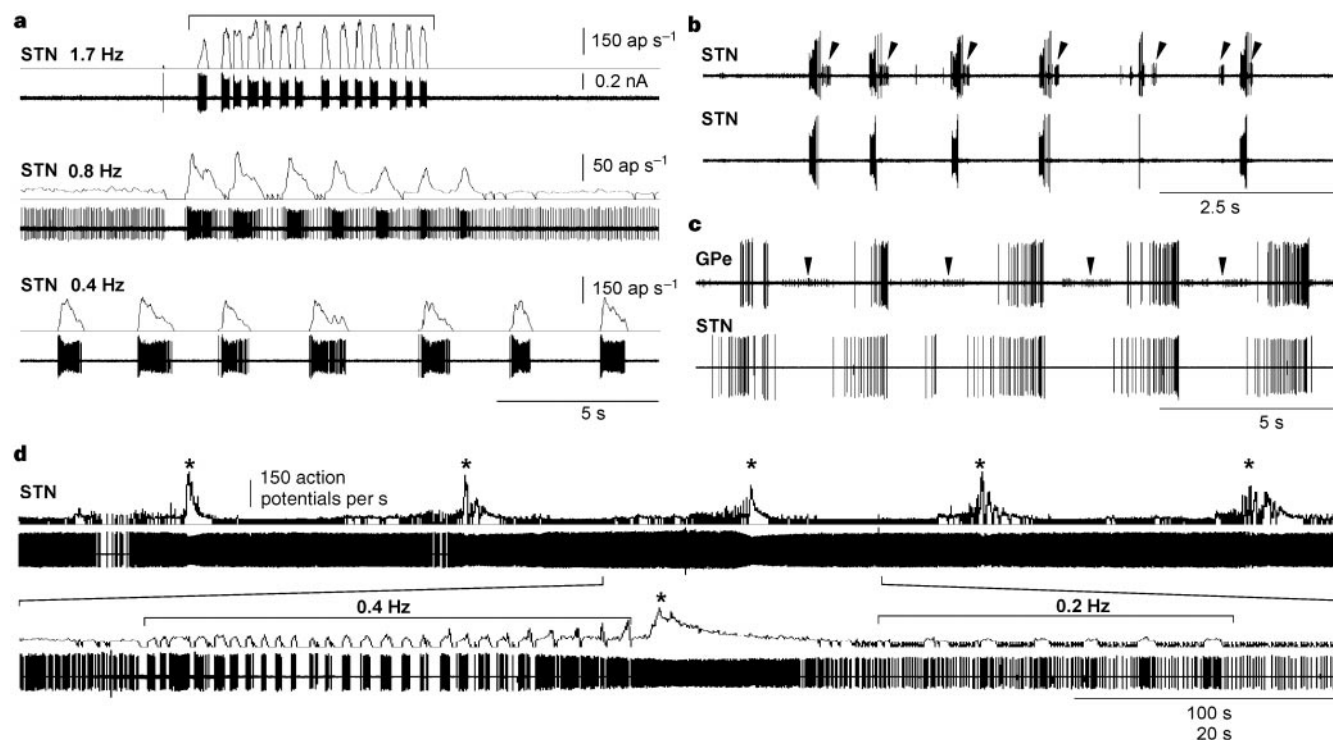


Figure 1 STN units display periods of oscillatory bursting, synchronized with other STN and GPe units. **a**, Oscillatory bursting periods with basic frequencies of 1.7, 0.8 and 0.4 Hz. Upper traces: instantaneous firing rates ($\tau = 0.1$ s). Lower traces: extracellular STN single-unit activity from different cultures. **b**, **c**, During synchronized bursting, STN units show stable phase relationships with other

STN or GPe units at 0° , close to 0° (**b**, arrowheads), or 180° (**c**, arrowheads). Simultaneous extracellular multi-unit recording with two electrodes. **d**, Synchronized bursting periods are grouped regularly into longer periods of recurrent neuronal activity (1–2 min). During periods of synchronized bursting (asterisks), shifts in intraburst frequencies occasionally occur (brackets).

main frequency was found by plotting f_0 against the corresponding total relative power in the $f_0 \pm \Delta f$ range for each correlation function (Fig. 2b). The resulting probability distribution of f_0 peaked at the previously obtained frequencies $f_{01} = 0.42 \pm 0.09$ and $f_{02} = 0.82 \pm 0.11$, but also revealed a third dominant frequency at $f_{03} = 1.87 \pm 0.21$ Hz (gaussian fits, mean $\pm$ s.d.; Fig. 2c) with an eight-times smaller relative power than that of f_{01} and f_{02} (0.0029 ± 0.0001 vs. 0.022 ± 0.001 ; group threshold, 1.0 Hz). All main frequencies were found for STN, GPe and type f_0 correlation function (Fig. 2b, analysis of variance (ANOVA)). STN neurons fired in distinct phase-relation to other STN and GPe neurons during burst firing. Phase-locked bursting was observed at all main frequencies (Fig. 2d, ANOVA), and phases of minimum and maximum ($\Phi_{\min}$, $\Phi_{\max}$) clustered mainly around (0° , $\pm 180^\circ$) or ($\pm 180^\circ$, 0°) (Fig. 2e). These results characterize phase-locked oscillatory bursting as a synchronized network activity within the STN–GPe system that stabilizes at three discrete harmonic frequencies during which neurons either burst together or alternate in bursting.

Although both STN and GPe neurons engage in burst activity, a quantitative comparison of STN–STN and STN–GPe cross-correlation functions showed that per single burst, STN units produced four times more spikes above expectation than GPe neurons (Fig. 2f).

The synchronized bursting was mainly confined to the low-frequency range. Using correlation analysis at higher time resolution (2 ms), oscillatory activity above 4 Hz was found in 12% of subthalamic (22/181) and 22% of pallidal (23/102) units. These oscillations were tightly correlated with a unit's average firing rate ($r_{\text{STN}} = 0.956$; $r_{\text{GPe}} = 0.998$) and were never found in the cross-correlation function (STN–STN, 0/73; STN–GPe, 0/76). Most units with these higher oscillations did not show synchronized oscillatory burst discharge (41/45). Thus, rhythmic firing at $f_0 > 4$ Hz probably reflected a neuron's intrinsic repetitive firing and not synchronized population activity.

In vivo, cortical stimulation produces complex sequences of excitation and inhibition in STN¹² and GPe¹³ that also involve the corticostriatopallidal pathway. To test whether cortical input is necessary for synchronized bursting, the cortical culture was acutely isolated by a cut along the corticostriatal border (Fig. 3A). Despite a more than 50% reduction in spike activity in the STN (Fig. 3D), the cortical lesion did not abolish synchronized oscillatory bursting ($n = 5$ cultures, Fig. 3A–C). However, it centred the f_0 distribution at f_{02} (Fig. 3E), indicating that f_{02} is the preferred oscillatory state in the absence of afferent cortical drive. This experiment shows that synchronized bursting in the STN and GPe system is very robust, which was further supported by the finding that the average intraburst interspike interval did not change for STN units following cortical lesion (Fig. 3F). Cortical lesion did not change firing rates in the GPe. As striatal activity could also be important in synchronized bursting, striatal and cortical inputs were interrupted by a cut placed at the striatum–STN border. Under these conditions, synchronized bursting was still present ($n = 2$, data not shown).

In vivo, the main inhibitory input to the STN originates from the GPe¹⁴, and the excitatory STN projects back to the GPe^{15,16}. This feedback system is crucial for the generation of synchronized bursts. After projections between the STN and GPe were interrupted by a cut along the STN–GPe border ($n = 4$ cultures), the synchronized bursting was abolished (Fig. 3G, H) and STN unit firing became regular, as indicated by a more than 70% decrease in the coefficient of variance ($P < 0.005$; Fig. 3I). Also, the average firing rates of STN neurons increased slightly (13 ± 2 Hz intact vs 16 ± 2 Hz lesioned), and most STN units (13/20) intrinsically oscillated ($r = 0.96$, f_0 vs average firing rate) with no correlation in the cross-correlation function (13/13). Within 1 h after the lesion, long-lasting and uncorrelated burst activity built up in the STN (Fig. 3J). Spontaneous activity was almost absent in the isolated GPe (data not shown). These results show that the GPe is crucial for the generation

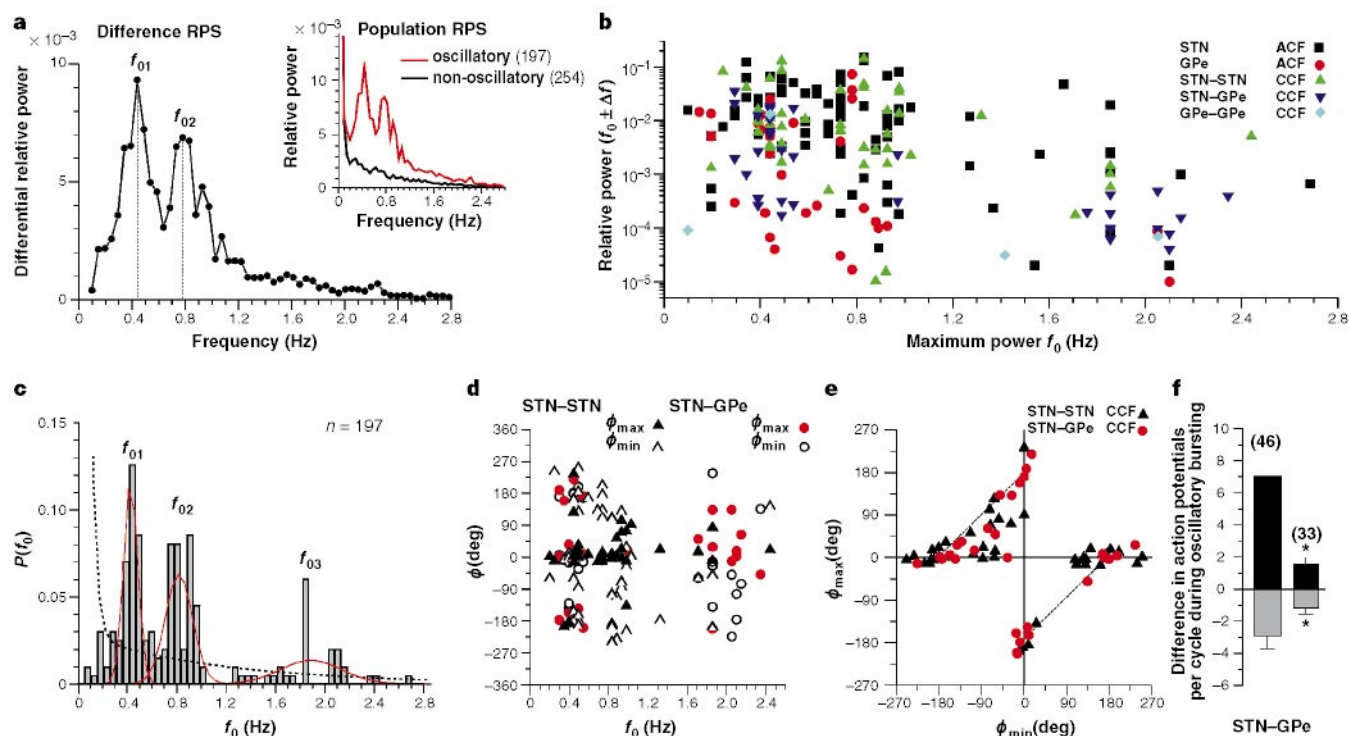


Figure 2 Quantitative analysis of synchronized oscillatory bursting in the STN and GPe. **a**, The difference relative power spectrum (RPS) reveals the tendency for bursting to occur at two main frequencies ($f_{01} = 0.44$ Hz, $f_{02} = 0.79$ Hz). Inset: average RPS from non-oscillatory (black) and oscillatory (red) auto- and cross-correlation functions (ACF and CCF, respectively). **b**, Scatter plot of f_0 versus relative power ($f_0 \pm \Delta f$) for each oscillatory correlation function and RPS. **c**, The probability distribution for f_0 [$P(f_0)$] obtained from **b** reveals three main population frequencies at $f_{01} = 0.42 \pm 0.09$, $f_{02} = 0.82 \pm 0.11$ and $f_{03} = 1.87 \pm 0.21$ (gaussian fit; red lines). Black dotted line indicates $P(f_0)$ for non-oscillatory cases (second-

order exponential decay fit). **d**, Phase-locked firing between STN and GPe units is independent of f_0 . **e**, Most units fire either synchronously at 0° or antiphasically at 180° with other STN and GPe units as revealed by the scatter plot of $(\phi_{\min}, \phi_{\max})$. Dashed lines indicate the expectation under the assumptions of symmetric burst-pause periods and equal distribution of all possible phases. **f**, STN units contribute significantly more spikes above chance ($\alpha_{0.995}$) than GPe units during one period of synchronized bursting ($P < 0.005$). Similarly, pauses lack significantly more spikes in STN neurons than in GPe neurons ($P < 0.05$).

of synchronized bursting and pallidal inputs are primarily responsible for the temporal organization of STN activity, rather than for the control of average firing rates of STN neurons.

Pharmacological studies support our model. Synchronized firing was blocked by bath application of the glutamate antagonist 6,7-dinitroquinoxaline-2,3-dione (DNQX) ($n = 3$ cultures, $n = 7$ neurons; Fig. 4a). Furthermore, STN burst activity elicited by local glutamate application to the STN initiated synchronized oscillatory bursting ($n = 5$; Fig. 4b). These results indicate that glutamatergic burst transmission from the STN to GPe is necessary for synchronized oscillatory bursting.

In addition, synchronized GABA (γ -aminobutyric acid) inputs to the STN, mimicked by local GABA application to the STN, elicited synchronized bursting that was preceded by a rebound burst in STN neurons ($n = 4$, Fig. 4c). This response profile indicates that synchronized inhibitory pallidal bursts could trigger synchronized oscillatory bursting by recruiting STN rebound bursts. The essential requirement of burst transmission between STN and GPe neurons for generating synchronized oscillatory activity was further supported by intracellular recordings from STN neurons combined with extracellular recordings in the GPe. Recordings were performed in cortex–STN–GPe triple cultures to exclude inhibitory inputs from the striatum. The data showed that pallidal bursts preceded hyperpolarizing, inhibitory potentials in STN neurons that were followed by rebound bursts ($n = 3$; Fig. 4d). These rebound bursts probably reflected intrinsic membrane properties of STN neurons, as rebound spiking could also be elicited by hyperpolarizing current injection into the cell bodies of cultured STN neurons (12/14; Fig. 4e)^{17,18}. Finally, in cortex–striatum–STN

cultures without GPe, spontaneous activity in the STN was characterized by long-lasting intermittent bursts (Fig. 4f) similar to those found in quadruple cultures after 1 h of acute GPe lesion.

The synchronized oscillatory bursting described here *in vitro* is very similar to activity patterns seen *in vivo* in the STN and GPe. In paralysed¹⁹ and anaesthetized rats (M. Bevan, personal communication), oscillatory burst discharge in the low-frequency range was found in the STN and GPe. In primates, most GPe neurons fire in periods of several hundred milliseconds of high-frequency discharge, interspaced with similarly long-lasting silent periods^{20–22}. These similarities at the neuronal activity level and the mature morphology of cortex–striatum–STN–GPe cultures¹¹ lead us to conclude that the synchronized oscillatory bursting described here captures a major activity pattern observed *in vivo*.

Our study identifies the STN–GPe system as a central pacemaker of the basal ganglia and should have far-reaching implications for basal ganglia function and dysfunction. First, in the culture system, the GPe was crucial for the temporal organization of STN activity, and GPe excitation was dictated by the STN. On the other hand, oscillatory bursts were less often encountered in the GPe than in the STN, and the burst strength of GPe units was less than that of STN units. This indicates that a few GPe neurons can have considerable control over the generation of synchronized oscillatory bursting within a small range of pallidal activity. As synchronized oscillatory burst discharge *per se* does not require cortical or striatal inputs, striatal inhibition of external pallidal neurons should allow for independent and very effective afferent control of the STN–GPe pacemaker.

Second, electrophysiological^{5,19,23,24} and *in situ* hybridization²⁵ results from animal models of Parkinson's disease have indicated

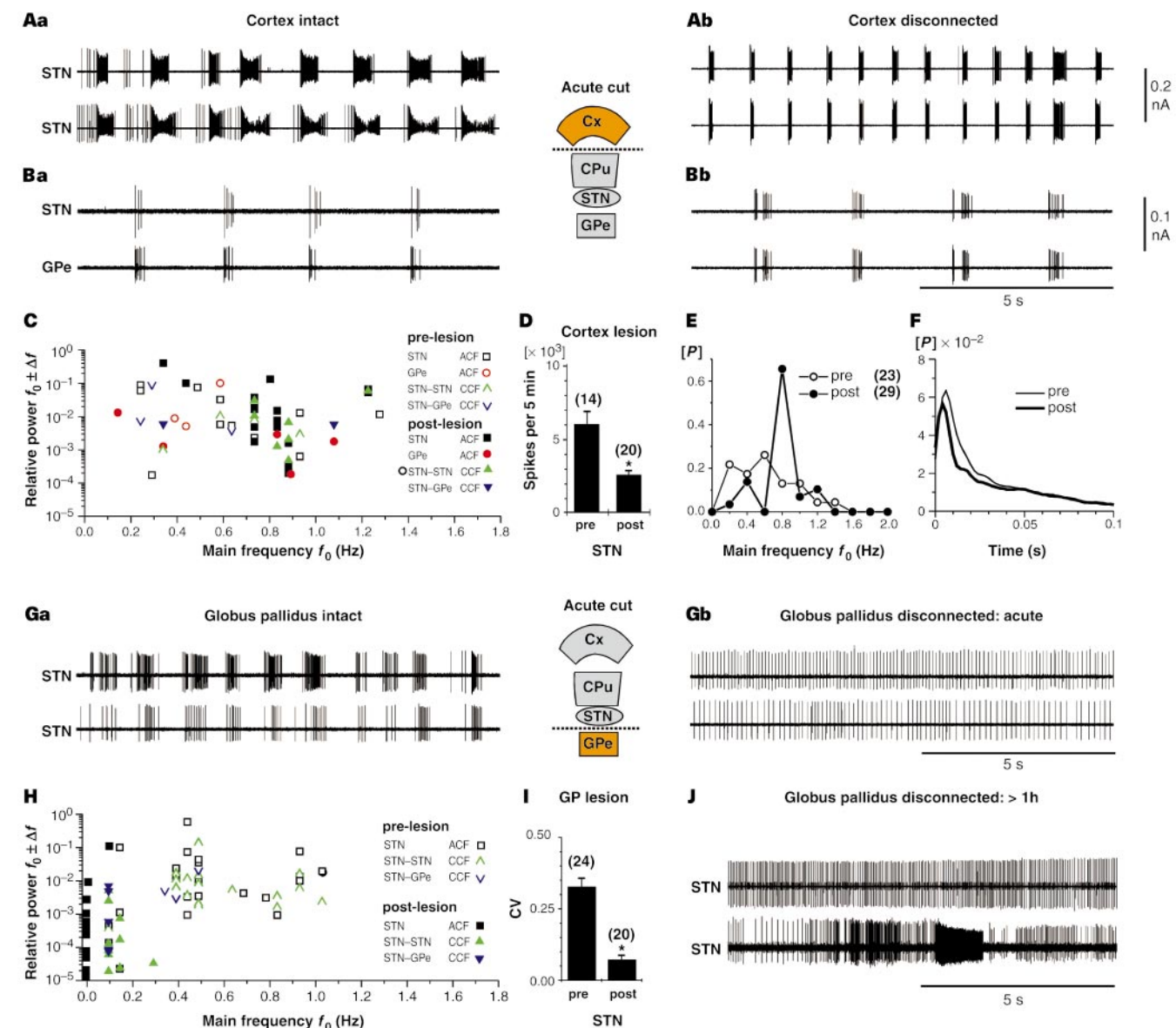


Figure 3 Synchronized oscillatory bursting does not depend on cortical input but on intact STN–GPe connectivity. **A, B**, The bursting does not depend on cortical inputs. Simultaneous recordings from different pairs of units in STN and GPe before (**Aa, Ba**) and after (**Ab, Bb**) cortical lesions. **C**, Scatter plot of f_0 and relative power in the $f_0 \pm \Delta f$ band before (pre-lesion, open symbols) and after (post-lesion, closed symbols) cortical inputs, synchronized bursting occurs predominantly at $f_0 = 0.8$ Hz. **F**, The average intraburst interspike interval histogram for STN units does not change after cortical lesion. **G**, Acute pallidal lesion abolishes

synchronized bursting in STN. Simultaneous recordings from different pairs of units before (**Ga**) and after (**Gb**) pallidal lesion. **H**, Scatter plot of f_0 and relative power in the $f_0 \pm \Delta f$ band before (pre-lesion, open symbols) and after (post-lesion, closed symbols) pallidal lesion. **I**, Reduction in coefficient of variance (CV) of STN firing after acute pallidal lesion. **J**, Simultaneous recording from STN units that display non-synchronized, non-oscillatory burst discharge after prolonged pallidal lesion (> 1 h). Cx, cortex; CPu, striatum.

that the temporal properties of STN activity^{2,5,25} and synchronization of neuronal activity in STN target nuclei^{9,10} might be important for basal ganglia dysfunction resulting from striatal dopamine deficiency. Our results provide a basic mechanism by which synchronized oscillatory activity in the STN–GPe pacemaker could be linked to dopamine deficiency in the striatum. We suggest that dopamine, by acting on the ‘indirect’ striatopallidal pathway³, controls the recruitment of STN and GPe neurons that become synchronized during pacemaker activity.

Finally, dopamine inputs were not present in the cortex–striatum–STN–GPe cultures. Thus, our *in vitro* system might more closely resemble a dopamine-deficient basal ganglia system rather than normal *in vivo* basal ganglia activity. This view is supported by the finding that dopamine depletion increases burst activity in the STN^{5,23,24}, external pallidum^{21,26} and basal ganglia

output nuclei^{21,27}. Nevertheless, the highly synchronized oscillatory firing might only be the simplest detectable mode of the STN–GPe pacemaker, given widespread synchronization in both nuclei and possible entrained activity in STN target nuclei. Under normal conditions, we would envisage pacemaker activity in multiple STN–GPe feedback loops that attain dynamic phase relationships with each other and are individually controlled by striatal inhibition. As the STN–GPe pacemaker is in the ‘indirect’ pathway of the basal ganglia^{3,15}, the activity in striatal neurons that constitute the ‘indirect’ pathway might be qualitatively different from striatal activity in the ‘direct’ pathway that has been directly linked to movement execution^{28,29}. The ‘indirect’ striatopallidal pathway would control temporal sequences of ‘windows’ through which activity in the ‘direct’ striatopallidal pathway is integrated by the output nuclei of the basal ganglia. □

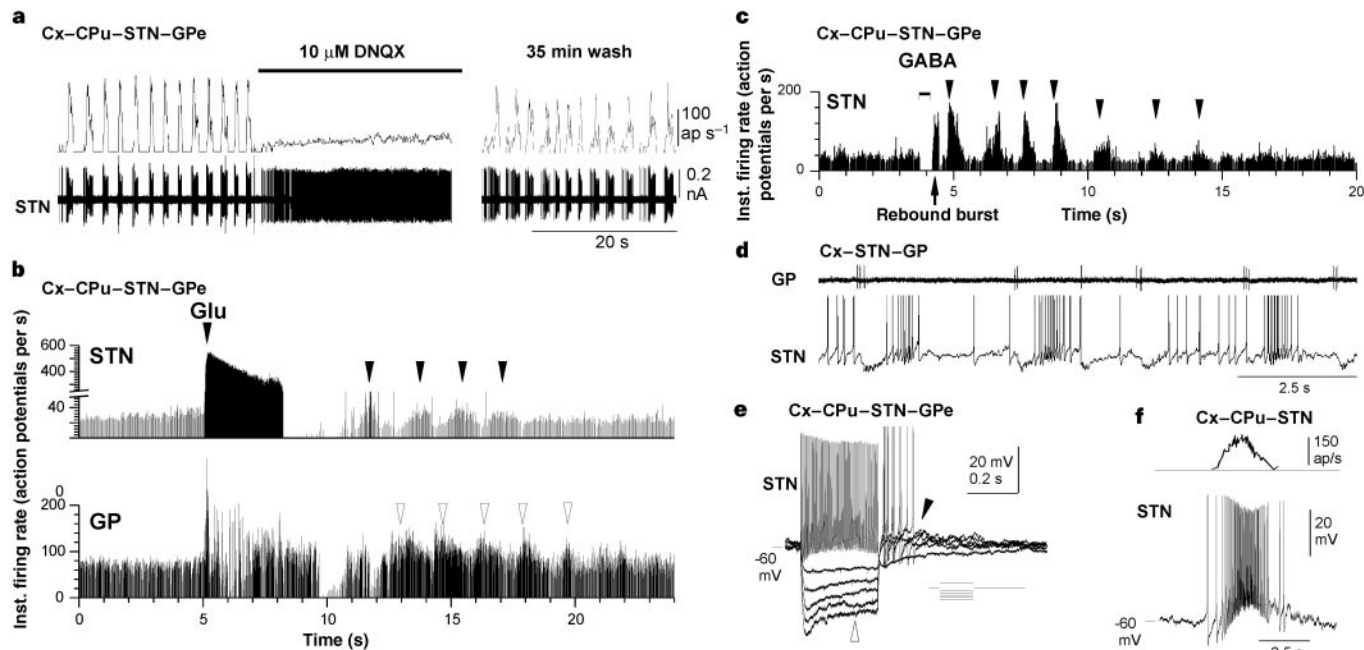


Figure 4 The mechanism of synchronized oscillatory burst generation in the STN-GPe pacemaker. **a**, Synchronized bursting is abolished by bath application of the glutamate antagonist DNQX. **b**, Synchronized oscillatory bursting can be elicited by local glutamate ejection (Glu) to the STN (200 ms, 1 mM). Instantaneous spike frequency plot for a simultaneously recorded STN and GPe unit. Note antiphasic bursting in STN and GPe (filled and open arrowheads). **c**, Local GABA application to the STN (200 ms, 1 mM) stops firing (bracket) and initiates oscillatory bursting (arrowheads) that is preceded by a rebound burst (arrow).

Methods

Preparation of cultures. From coronal sections of rat brains at postnatal day 0–2 (Sprague-Dawley Harlan), dissected areas containing cortical, striatal, subthalamic and pallidal tissue were cultured in sequential order using a modified 'rollertube' technique¹¹. The STN and GPe could be clearly identified in the mature culture during recording.

Electrophysiology. The extracellular solution contained (in mM) 126 NaCl, 0.3 NaH₂PO₄, 2.5 KCl, 0.3 KH₂PO₄, 1.6 CaCl₂, 1.0 MgCl₂, 0.4 MgSO₄, 26.2 NaHCO₃, and 11 D-glucose, saturated with 95% O₂ and 5% CO₂ (artificial cerebrospinal fluid; ACSF) warmed to 36.5 ± 1 °C. Extracellular recordings were obtained with patch electrodes filled with ACSF in current-follower mode, positioned under visual control. Intracellular recordings were obtained with sharp microelectrodes (110–150 MΩ) containing 2 M potassium acetate and 2% neurobiotin (Vector). Signals were recorded in Spike2 (Cambridge Electronic Design Ltd) and spike discharge was detected off-line using logic threshold operations or spike-template matching. Extracellular spike amplitudes and waveforms did vary during synchronized bursting (Figs 1, 3). Thus, to ensure proper single unit isolation, most units were isolated using a threshold operation and large signal-to-noise ratio (~6–10:1). Our data set mainly comprised single-unit activity.

Data analysis. Correlation functions were calculated for each unit at 2 and 20 ms time resolution (1,024 bins; ±1.024 s and ±10.24 s, respectively), smoothed (Fourier filter, quadratic function, cut off at 10 and 100 Hz), and normalized to average firing rates with confidence intervals ($\alpha_{0.005}$, $\alpha_{0.995}$) based on Poisson statistics using mean firing rates³⁰. The relative power spectrum (RPS) was obtained by fast Fourier transformation of correlation functions (Mathematica) and normalization to the DC component. A single neuron or pair of neurons were considered oscillatory if the correlation function crossed at least twice the upper and lower confidence levels of multiple periods corresponding to the inverse of the main frequency $f_0 \pm 2\Delta f$ taken from the RPS. A peak detection algorithm was used to find the maximum peak in each RPS that defined the main frequency $f_0 > 0.08$ Hz for each oscillatory and non-oscillatory correlation function. We obtained the difference population RPS by

subtracting the average RPS of all non-oscillatory units from that of oscillatory units. The phases of maximum ($\phi_{\max}$) and minimum ($\phi_{\min}$) indicate the time-to-maximum or time-to-minimum normalized to f_0^{-1} in crosscorrelation functions. Relative burst strength was calculated by integrating the area above or below expectation ($\alpha_{0.005}$, $\alpha_{0.995}$) during the burst oscillation period f_0^{-1} around time zero in STN-STN and STN-GPe crosscorrelation functions. Data are expressed as mean ± s.e.m. For statistical analysis, a factorial ANOVA with a *post hoc* Student–Newman–Keuls test (Statview) or Mann–Whitney *U*-test were used (significance level, $P < 0.05$). Correlation was estimated by linear regression analysis.

Pharmacology and anatomical reconstruction. GABA (1 mM), glutamate (1 mM; all Sigma), and DNQX (RBI) were dissolved in ACSF and either bath-applied or locally pressure-ejected close to the cell body of the neuron from which was being recorded. Acute cortical and pallidal lesions were performed with a micro-scalpel attached to a micromanipulator under visual guidance, after which new recordings were established. Lesion success was controlled for by the disappearance of evoked responses in the STN upon cortical or pallidal stimulation (tungsten stimulation electrode, 50 μs duration, 80–120 μA current amplitude). For each culture, we used parvalbumin immunohistochemistry¹¹ to verify recording and lesion positions anatomically.

Received 22 March; accepted 1 June 1999.

1. Matsumura, M., Kojima, J., Gardiner, T. W. & Hikosaka, O. Visual and oculomotor functions of the monkey subthalamic nucleus. *J. Neurophysiol.* **67**, 1615–1632 (1992).
2. Wichmann, T., Bergman, H. & DeLong, M. R. The primate subthalamic nucleus. I. Functional properties in intact animals. *J. Neurophysiol.* **72**, 494–506 (1994).
3. Albin, R. L., Young, A. B. & Penney, J. B. The functional anatomy of basal ganglia disorders. *Trends Neurosci.* **12**, 366–375 (1989).
4. Wichmann, T. & DeLong, M. R. Functional and pathophysiological models of the basal ganglia. *Curr. Opin. Neurobiol.* **6**, 751–758 (1989).
5. Bergman, H., Wichmann, T., Karmon, B. & DeLong, M. R. The primate subthalamic nucleus. II. Neuronal activity in the MPTP model of parkinsonism. *J. Neurophysiol.* **72**, 507–520 (1994).
6. Bergman, H., Wichmann, T. & DeLong, M. R. Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science* **249**, 1436–1438 (1990).
7. Limousin, P. *et al.* Electrical stimulation of the subthalamic nucleus in advanced Parkinson's disease. *N. Engl. J. Med.* **339**, 1105–1111 (1998).
8. Rodriguez, M. C. *et al.* The subthalamic nucleus and tremor in Parkinson's disease. *Mov. Disord.* **13**, 111–118 (1998).

9. Nini, A., Feingold, A., Slovins, H. & Bergman, H. Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the MPTP model of Parkinson. *J. Neurophysiol.* **74**, 1800–1805 (1995).
10. Hurtado, J. M., Gray, C. M., Tamas, L. B. & Sigvardt, K. A. Dynamics of tremor-related oscillations in the human globus pallidus: A single case study. *Proc. Natl Acad. Sci. USA* **96**, 1674–1679 (1999).
11. Plenz, D., Herrera-Marschitz, M. & Kitai, S. T. Morphological organization of the globus pallidus–subthalamic nucleus system studied in organotypic cultures. *J. Comp. Neurol.* **397**, 437–457 (1998).
12. Kitai, S. T. & Deniau, J. M. Cortical inputs to the subthalamus: intracellular analysis. *Brain Res.* **214**, 411–415 (1981).
13. Kita, H. Responses of globus pallidus neurons to cortical stimulation: intracellular study in the rat. *Brain Res.* **589**, 84–90 (1992).
14. Kita, H., Chang, H. T. & Kitai, S. T. Pallidal inputs to subthalamus: intracellular analysis. *Brain Res.* **264**, 255–265 (1983).
15. Parent, A. & Hazrati, L. N. Functional anatomy of the basal ganglia. II. The place of the subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Res. Rev.* **20**, 128–154 (1995).
16. Shink, E., Bevan, M. D., Bolam, J. P. & Smith, Y. The subthalamic nucleus and the external pallidum: two tightly interconnected structures that control the output of the basal ganglia in the monkey. *Neuroscience* **73**, 335–357 (1996).
17. Nakanishi, H., Kita, H. & Kitai, S. T. Electrical membrane properties of rat subthalamic neurons in an *in vitro* slice preparation. *Brain Res.* **437**, 35–44 (1987).
18. Beurrier, C., Congar, P., Bioulac, B. & Hammond, C. Subthalamic nucleus neurons switch from single-spike activity to burst firing mode. *J. Neurosci.* **15**, 599–609 (1999).
19. Ruskin, D. N. *et al.* Multisecond oscillations in firing rate in the basal ganglia: Robust modulation by dopamine receptor activation and anaesthesia. *J. Neurophysiol.* **81**, 2046–2055 (1999).
20. DeLong, M. R. Activity of pallidal neurons during movement. *J. Neurophysiol.* **34**, 414–427 (1971).
21. Filion, M. Effects of the interruption of the nigrostriatal pathway and of dopaminergic agents on the spontaneous activity of globus pallidus neurons in the awake monkey. *Brain Res.* **178**, 425–441 (1978).
22. Aldridge, J. W. & Gilman, S. The temporal structure of spike trains in the primate basal ganglia: Afferent regulation of bursting demonstrated with precentral cerebral cortical ablation. *Brain Res.* **543**, 123–138 (1991).
23. Hollerman, J. R. & Grace, A. A. Subthalamic nucleus cell firing in the 6-OHDA-treated rat: basal activity and response to haloperidol. *Brain Res.* **590**, 291–299 (1992).
24. Hassani, O.-K., Mouroux, M. & Féger, J. Increased subthalamic neuronal activity after nigral dopaminergic lesion independent of disinhibition via the globus pallidus. *Neurosci.* **72**, 105–115 (1996).
25. Chesselet, M. F. & Delfs, J. M. Basal ganglia and movement disorders: an update. *Trends Neurosci.* **10**, 417–422 (1996).
26. Pan, H. S. & Walters, J. R. Unilateral lesion of the nigrostriatal pathway decreases the firing rate and alters the firing pattern of globus pallidus neurons in the rat. *Synapse* **2**, 650–656 (1988).
27. Miller, W. C. & DeLong, M. R. In *The Basal Ganglia II* (eds Carpenter, M. B. & Jayaraman, A.) 415–427 (Plenum, New York, 1987).
28. Wurtz, R. H. & Hikosaka, O. Role of the basal ganglia in the initiation of saccadic eye movements. *Prog. Brain Res.* **64**, 175–190 (1986).
29. Chavali, G. & Deniau, J. M. Disinhibition as a basic process in the expression of striatal functions. *Trends Neurosci.* **13**, 277–280 (1990).
30. Abeles, M. Quantification, smoothing, and confidence limits for single-units' histograms. *J. Neurosci. Methods* **5**, 317–325 (1982).

Acknowledgements. We thank B. Teng for technical assistance with the preparation of cultures and for immunohistochemistry, and H. Steiner for discussions. This work was supported by grants from the National Parkinson Foundation (to D.P.) and the NINDS (to D.P. and S.T.K.).

Correspondence and requests for materials should be addressed to D.P. (e-mail: dplenz@codon.nih.gov).

Continued RAG expression in late stages of B cell development and no apparent re-induction after immunization

Wong Yu^{†‡}, Hitoshi Nagaoka^{†‡}, Mila Jankovic[§], Ziva Misulovin^{†‡}, Heikyung Suh^{†‡}, Antonius Rolink^{||}, Fritz Melchers^{||}, Eric Meffre^{†‡} & Michel C. Nussenzweig^{†‡}

^{*}Laboratory of Molecular Immunology, and the [†]Howard Hughes Medical Institute, the Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

[§]Institute for Molecular Genetics and Genetic Engineering, Vojvode Stepe 444A, Belgrade, Yugoslavia

^{||}Basel Institute for Immunology, Grenzacherstrasse 487, Postfach, CH-4005 Basel, Switzerland

[†]These authors contributed equally to this work.

Models of B-cell development in the immune system suggest that only those immature B cells in the bone marrow that undergo receptor editing express *V(D)J*-recombination-activating genes (RAGs)^{1–3}. Here we investigate the regulation of RAG expression in transgenic mice carrying a bacterial artificial chromosome that encodes a green fluorescent protein reporter instead of RAG2

(ref. 4). We find that the reporter is expressed in all immature B cells in the bone marrow and spleen. Endogenous RAG messenger RNA is expressed in immature B cells in bone marrow and spleen and decreases by two orders of magnitude as they acquire higher levels of surface immunoglobulin M (IgM). Once RAG expression is stopped it is not re-induced during immune responses. Our findings may help to reconcile a series of apparently contradictory observations, and suggest a new model for the mechanisms that regulate allelic exclusion, receptor editing and tolerance.

To investigate the regulation of RAG expression, we produced transgenic mice that carry bacterial artificial chromosomes (BACs) modified by homologous recombination to encode a green fluorescent protein (GFP) reporter instead of RAG2 (ref. 4) (Fig. 1a). Three independently derived strains of NG-BAC transgenic mice were analysed and found to be identical in terms of GFP expression.

To determine whether GFP expression in B cells in NG-BAC transgenic mice reflects documented patterns of RAG expression, we examined bone-marrow cells by using multi-parameter flow cytometry. B220⁺CD43[−] non-B cells and early precursors not known to express RAGs did not express GFP³ (Fig. 1b, G1). In contrast, a high level of GFP expression was found in B220⁺CD43⁺ pro-B cells and a lower level was found in B220^{low}CD43[−] pre-B cells (Fig. 1b, G2 and G3, and Fig. 1c). Consistent with previous studies of RAG messenger RNA and protein expression in pre-B cells³, GFP expression was lower in large B220^{low}CD43[−] IgM[−] pre-B cells than in small pre-B cells (Fig. 1c). The finding that GFP fluorescence reflects RAG2 protein expression in small and large pre-B cells indicates that GFP is appropriately modulated in these cells despite the absence in GFP of amino-acid sequences that influence RAG2 stability⁵. Finally, mature B220^{high}CD43[−] B cells that re-circulate through the bone marrow did not express significant levels of GFP (Fig. 1b, G4). We conclude that the expression of GFP, like that of RAG2, is specifically induced in pro-B cells, decreases with the pre-B-cell transition, varies between large and small pre-B cells, and is turned off in mature B cells.

Immature B cells can be distinguished from earlier precursors in the bone marrow and from mature re-circulating bone marrow B cells because they are Igκ⁺B220^{low}HSA^{high} (ref. 6). We found that all Igκ⁺ and B220^{low} or HSA^{high} bone-marrow B cells are also GFP⁺, whereas GFP⁺Igκ⁺ B cells display a B220^{high} and HSA^{low} phenotype consistent with mature re-circulating B cells (Fig. 1d). Thus, immature B cells in the bone marrow uniformly express the RAG2–GFP reporter but at a lower level than pro- and pre-B cells (Fig. 1c, d).

To determine whether immature B cells continue to express the GFP indicator after leaving the bone marrow, we analysed splenic B cells from NG-BAC transgenic mice. We found that 5–30% of all B220⁺ cells or 2–12% of all lymphocytes are GFP⁺ (Fig. 2a). Mice that are 3–4 weeks old, have higher proportions of GFP⁺ B cells than older mice 6–24 weeks old (not shown). The GFP⁺ B cells in the spleen are IgM⁺ and either Igκ⁺ or Igλ⁺, so they are not contaminating B-cell precursors (Fig. 2a and data not shown). To determine whether the GFP⁺ B cells in the spleen are transitional B cells, we analysed spleen cells from NG-BAC transgenic mice using a combination of anti-B220, -HSA and 493 antibodies^{6,7}. The antigen recognized by the 493 antibody is expressed on transitional but not mature B cells in the spleen⁷. Transitional B cells (B220^{low}HSA^{high}493⁺) are uniformly GFP⁺, whereas all mature B cells (B220^{high}HSA^{low}493[−]) are GFP[−] (Fig. 2b, c).

To determine whether GFP expression in immature and transitional B cells correlates with endogenous RAG expression we assayed for these mRNAs in B cells from non-transgenic mice. Consistent with our observations in GFP transgenic mice, mature B cells from wild-type mice do not express RAG1 or RAG2 mRNA or RAG2 protein (Fig. 3a, b). In contrast, immature B cells from bone marrow and spleen contain RAG1 and RAG2 but the level of RAG mRNA in these cells decreases by two orders of magnitude as they

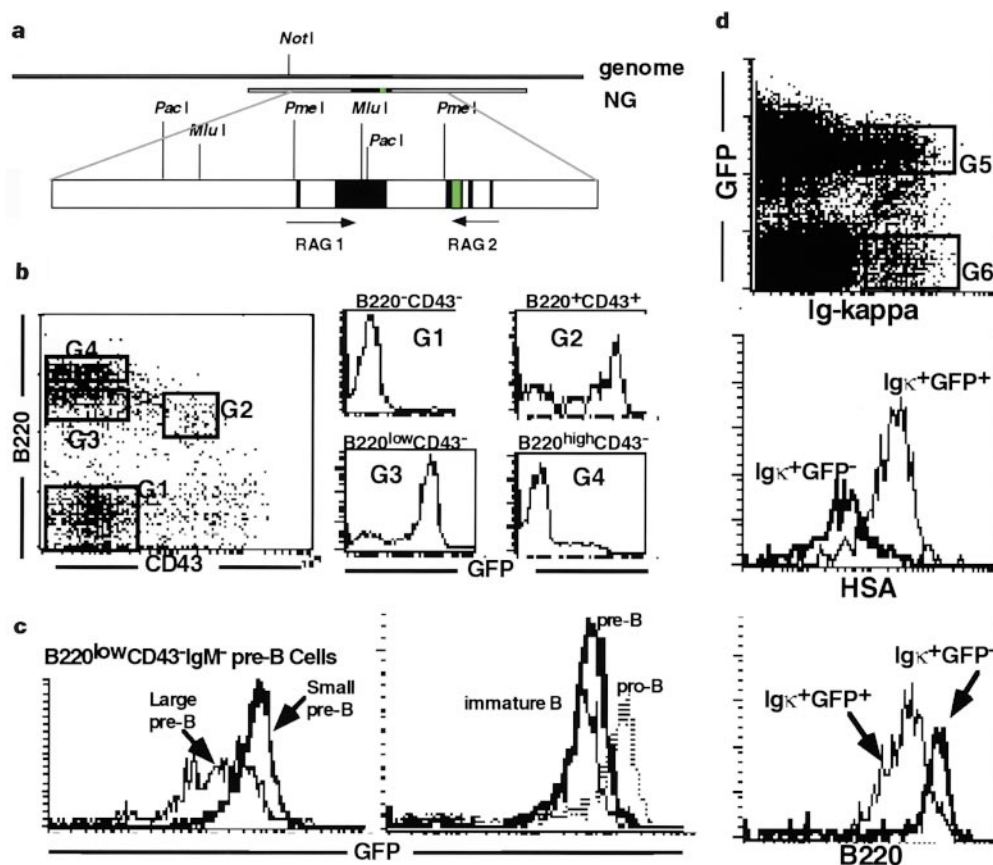


Figure 1 GFP expression in B cells from NG-BAC transgenic mice 6–8 weeks old. **a**, The RAG locus and GFP-modified NG-BAC; 55 kilobases (kb) of genomic DNA are found upstream of RAG1, and 85 kb are found upstream of RAG2. GFP (green) was inserted at the ATG of RAG2. **b**, GFP expression in bone-marrow cells electronically gated on the basis of B220 and CD43 staining, as indicated on the

dot plot. **c**, GFP expression in B220^{low}CD43⁻IgM⁻ bone-marrow pre-B cells separated on the basis of size by forward and side scatter parameters, and in B220^{low}CD43⁻IgM⁻ pro-B cells, B220^{low}CD43⁻IgM⁻ pre-B cells and B220^{low}CD43⁻IgM⁻ immature B cells. **d**, B220 or HSA expression on bone-marrow cells gated by Igκ or GFP expression, as indicated.

mature and acquire higher levels of surface IgM (Fig. 3a). There is three-to-five-fold less RAG mRNA in immature B cells with low levels of surface IgM than in pre-B cells, and a further five-to-twenty-five-fold drop when these cells express high levels of IgM. This gradual decrease in RAG mRNA is also seen in the immature B-cell compartment in the spleen but is not quantitatively reflected in decreased GFP fluorescence. GFP degradation lags behind RAG mRNA in immature and transitional B cells and GFP expression is therefore an indicator of recent RAG expression but cannot be used as a quantitative index of RAG mRNA levels. As RAG2 protein also persists longer than RAG mRNA³ in immature B cells, we measured RAG2 in immature B cells in spleen by immunoprecipitation and western blotting. Purified GFP⁺ and GFP⁻ splenic B cells were compared with B220⁺IgM⁻ B cell precursors from bone marrow. GFP⁺ cells from spleen contain RAG2 protein while GFP⁻ splenic B cells do not (Fig. 3b). Consistent with the drop in GFP fluorescence between B-cell precursors and transitional B cells (Fig. 1c, GFP levels decrease 1 log from pro-B to pre-B to immature B), RAG2 protein is 15-fold lower in immature spleen B cells than in B220⁺IgM⁻ B-cell precursors (Fig. 3b). It remains possible that endogenous RAGs differ from the indicator and are expressed at high levels in a subgroup of the GFP⁺B220^{low}IgM^{med/high} B cells in bone marrow and spleen. We have been unable to detect such a high expressing subgroup by *in situ* hybridization in purified IgM^{med/high} transitional B cells. We conclude that GFP⁺B220^{low}IgM^{low} B cells in the spleen express RAG2 protein in addition to RAG2 mRNA. However, the precise origin of the RAG-expressing IgM^{low} B cells in the spleen remains to be determined.

To determine whether RAG expression in B cells in spleen correlates with V(D)J recombination, we measured recombination signal sequence-specific DNA breaks in purified GFP⁺ and GFP⁻ splenic B cells. GFP⁺ B cells from spleen showed recombination and sequence breaks, whereas GFP⁻ mature B cells from spleen did not (Fig. 3c). Thus, only GFP⁺ splenic B cells show intermediates of V(D)J recombination, and GFP expression in peripheral B cells cannot be simply due to delayed degradation of GFP, because: GFP and RAGs are not found in peripheral T cells even though GFP is highly expressed in thymocytes (Fig. 2a and data not shown); endogenous RAG mRNA is present in immature B cells (Fig. 3a); endogenous RAG2 protein is found in B cells in spleen (Fig. 3b); and GFP expression is rapidly modulated in large and small pre-B cells in the bone marrow, and by receptor crosslinking *in vitro* (Fig. 1 and data not shown).

How is RAG expression in spleen B cells related to recent reports of RAG expression in germinal centre B cells? When RAG protein expression was detected by immunohistochemistry in germinal centre B cells and in spleen B cells treated with lipopolysaccharide (LPS) and interleukin-4 (IL-4), it was assumed that this was due to re-induction^{8,9}. However, the immunohistochemistry and subsequent functional experiments were not in complete agreement, as: V(D)J recombination was seen early in immune responses before germinal centre formation^{9,10}; not all germinal centres contained B cells that expressed the V-preB protein, which is found in B cells undergoing light chain recombination¹¹; and B cells outside germinal centres were found to express V-preB¹¹. Thus, the nature of the cells undergoing secondary recombination in peripheral lymphoid organs was not well defined.

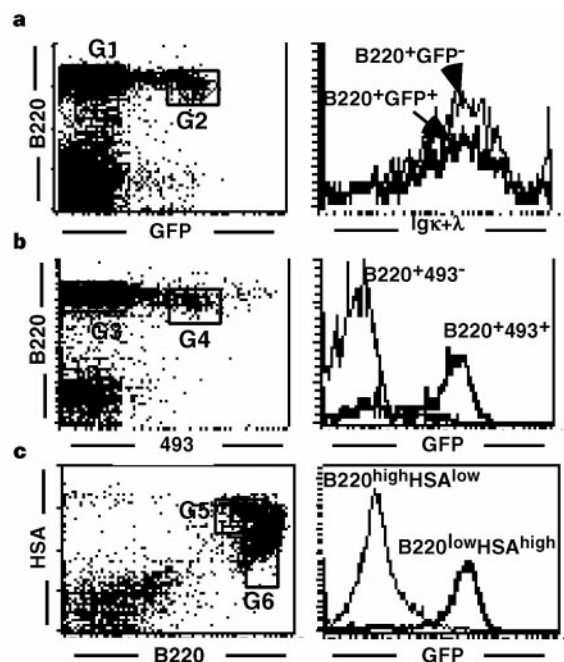


Figure 2 RAG expression in NG-BAC transgenic spleen cells. **a**, Igκ and Igλ expression on B220⁺GFP⁺ spleen B cells electronically gated as indicated (G2 and G1). **b**, GFP expression in B220^{low}493⁺ transitional B cells and B220^{high}493⁺ mature

B cells gated as indicated (G4 and G3). **c**, GFP expression in B220^{low}HSA^{high} transitional B cells and B220^{high}HSA^{low} mature B cells gated as indicated (G5 and G6).

To determine whether GFP⁺ B cells can form germinal centres, we performed immunization experiments in RAG1^{-/-} mice reconstituted by adoptive transfer of purified carrier primed T cells and either GFP⁺ or GFP⁻ splenic B cells. Serum samples and tissues were tested for specific anti-4-hydroxy-3-nitrophenyl acetyl (anti-NP)

antibody responses and the formation of germinal centres, respectively. Anti-NP IgM and IgG1 antibody responses and germinal centres were found in RAG1^{-/-} recipients of either GFP⁺ or GFP⁻ B cells 12 days after immunization (Fig. 4a, b). In contrast, no such response was found in control RAG1^{-/-} mice or in RAG1^{-/-} mice

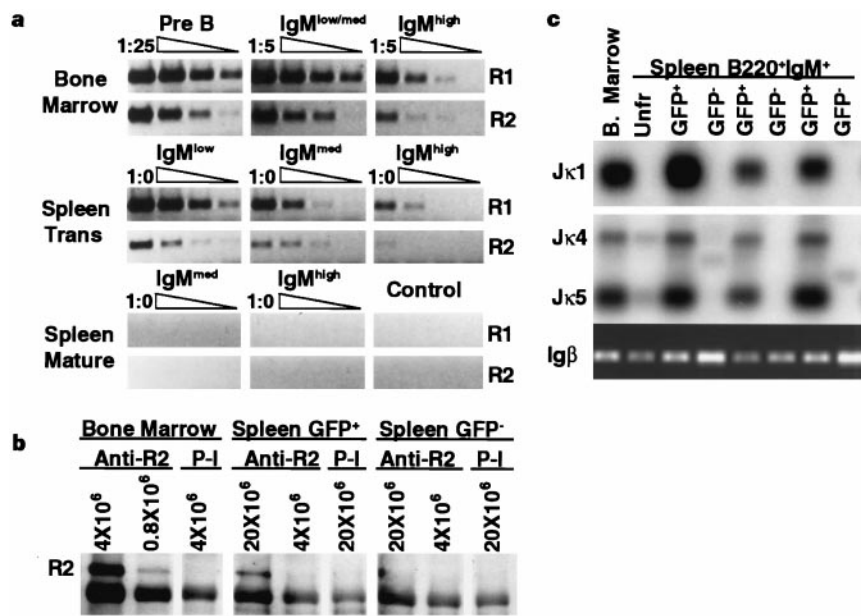


Figure 3 RAG2 expression and RSS breaks in purified spleen B cells. **a**, RAG1 (R1) or RAG2 (R2) mRNA expression in sorted populations of wild-type bone marrow and peripheral B cells detected by RT-PCR. Pre-B cells are IgM⁺B220^{low}CD43⁻; immature B220^{low} B cells were sorted into IgM^{low/med} and IgM^{high} groups. Transitional (Trans, B220^{low}HSA^{high}CD43⁻) and mature (B220^{high}HSA^{low}CD43⁻) B cells were further sorted into IgM^{low}, IgM^{med} and IgM^{high} groups. Serial five-fold dilutions of cDNA normalized by Igβ PCR and cell number are shown. The first dilutions of the pre-B, IgM^{low/med} and IgM^{high} bone marrow groups are 1:25, 1:5 and 1:5 respectively. The first dilution of all spleen samples is 1:0. No cDNA was

added to the negative controls. **b**, IP western blot for RAG2. The number of cells for each immunoprecipitation (polyclonal rabbit anti-RAG2 (Anti-R2) or pre-immune serum (P-I)) is indicated. Bone-marrow cells were depleted of Igκ⁺ and Igλ⁺ cells and enriched for B220⁺ cells. B220⁺GFP⁺ or GFP⁻ cells were purified from spleen by cell sorting. The lower band in all lanes is rabbit IgG heavy chain, which crossreacts with the secondary antibody. **c**, RSS breaks in purified spleen B cells. Spleen B cells were separated into GFP⁺ and GFP⁻ populations by fluorescence-activated cell sorting³⁰. PCR for Igβ is a DNA loading control.

that received only carrier primed T cells (Fig. 4a, b). We conclude that both GFP-expressing and GFP-non-expressing B cells can be recruited to germinal centres, and that both types of B cell can respond to T-cell-dependent antigens *in vivo*.

To determine whether mature B cells can be induced to express RAGs during immune responses, we performed additional adoptive transfer experiments. RAG1^{-/-} mice were immunized one day after reconstitution with carrier primed T cells and either GFP⁺ or GFP⁻ B cells and their splenic B cells were analysed for GFP expression 5, 8 and 13 days after transfer. We found that GFP⁻ B cells did not become GFP⁺ during the immune response, and GFP⁺ B cells became GFP⁻. Only 15% of the initially GFP⁺ B cells remained GFP⁺ on day 5 after transfer and, by day 13, 99% of the GFP⁺ B cells had become GFP⁻ (Fig. 4c). To verify that endogenous RAGs are not induced in mature B cells during immune responses, we compared RAG mRNA expression by RT-PCR in sorted B220⁺GL7⁺ cells from immunized control and reconstituted RAG1^{-/-} mice. As previously reported, B220⁺GL7⁺ cells from immunized wild-type mice express RAG1 mRNA⁸. However, there was no RAG expression in B220⁺GL7⁺ B cells purified from RAG1^{-/-} mice that had been

immunized after adoptive transfer of purified GFP⁻ B cells and carrier primed T cells (Fig. 4d, e). We conclude that, once RAG transcription ceases, it is not re-induced during an immune response *in vivo*.

To determine whether LPS and IL-4 together induce GFP indicator expression *in vitro*, we cultured purified GFP⁺ and GFP⁻ spleen B cells with LPS in the presence or absence of IL-4. GFP⁻ cells remained GFP⁻ even when treated with LPS and IL-4, and GFP⁺ B cells lost indicator expression with time just as they had done *in vivo* (Fig. 4c, f). However, GFP⁺ cells treated with LPS and IL-4 showed a higher percentage of GFP expression than GFP⁺ cells treated with LPS alone. This small relative increase in GFP expression may account for the apparent induction previously reported⁹.

What is the function, if any, of continued low-level RAG expression in the periphery? Immature B cells persist in the spleen for a few days and only a fraction of them are selected into the long-lived B-cell compartment^{6,7,12}. This developmental transition coincides with the cessation of RAG expression and seems to be modulated by the BCR. For example, the number of immature B cells varies in different strains of BCR transgenic mice and was

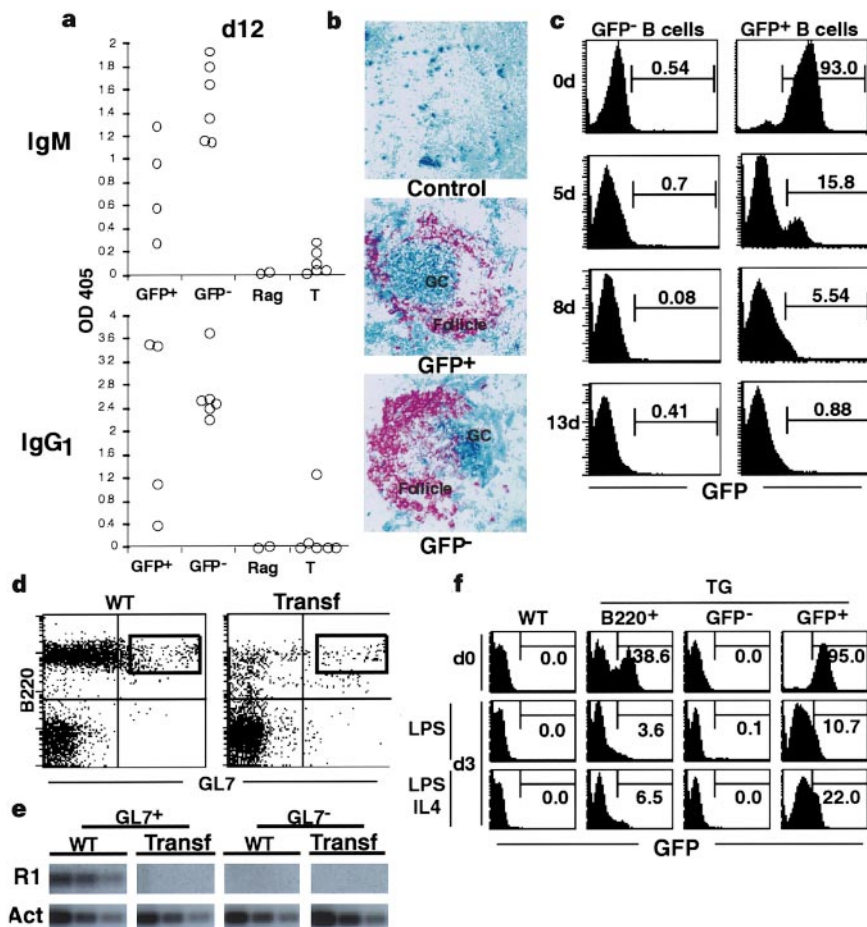


Figure 4 Immune response to NP-CGG by GFP⁺ and GFP⁻ spleen B cells. **a**, Anti-NP IgM and IgG1 responses in serum samples from RAG1^{-/-} mice reconstituted with carrier primed T cells and either GFP⁺ or GFP⁻ splenic B cells, as measured by ELISA. Each data point represents serum from an individual mouse taken on day 12 after immunization. **b**, Splenic germinal centres visualized by PNA (blue) and IgM (brown) staining. RAG1^{-/-} mice reconstituted with carrier primed T cells and either GFP⁺ or GFP⁻ spleen B cells 12 days after immunization. Controls were RAG1^{-/-} mice that did not receive B or T cells. **c**, GFP expression in B220⁺ spleen B cells from immunized mice on the indicated days after cell transfer. The percentage of GFP⁺ B cells is shown. Specific immune responses were not found in unimmunized mice. **d**, GL7 expression in B cells 12 days after

immunization. Gates used for cell sorting are shown. **e**, RAG1 expression by RT-PCR in sorted B220⁺GL7⁺ B cells from wild-type control and adoptively transferred recipients of carrier primed T cells and GFP⁻ spleen B cells. β -Actin is the mRNA loading control. Serial threefold dilutions of the cDNAs are shown. **f**, Addition of IL-4 prolongs GFP-RAG2 expression *in vitro*. Histograms show GFP intensity on B220⁺ live cells. Sorted GFP⁺/B220⁺ and GFP⁻/B220⁺ or unseparated NG and control splenocytes were cultured for 3 days in LPS, with or without IL4. WT, wild type; TG, transgenic splenocytes; GFP⁺, sorted B220⁺/GFP⁺ splenocytes; GFP⁻, sorted B220⁺/GFP⁻ splenocytes; Transf, adoptively transferred RAG1^{-/-} mouse; Act, β -actin.

increased in the spleen of V_HB1-8/Vk3-83 replacement mice that showed high levels of secondary recombination *in vitro*¹³. Although the nature of the selection from immature to mature compartments is not well defined, there is evidence for both positive and negative selection^{12,14,15}. Therefore, the selective pressures in the periphery may be similar to those in the immature B-cell compartment in the bone marrow. RAG expression in the IgM^{low} immature B cells in the bone marrow mediates editing and repertoire diversification^{1,2,16}, and these may also be the principal functions of RAG in IgM^{low} B cells in the spleen. RAG expression in peripheral B cells has been proposed as a mechanism for the diversification of immune responses in germinal centres^{8–11,17}. However, the results presented here limit this mechanism to a small subset of IgM^{low} B cells in the spleen that may enter germinal centres before they have turned off RAG expression.

The observation that IgM^{low} B cells in spleen and bone marrow continue to express RAGs means that B-cell development and light-chain allelic exclusion can be re-evaluated as processes with strong parallels to T-cell development and T-cell receptor (TCR)- α allelic exclusion (reviewed in ref. 18). IgM^{low} B cells in spleen and bone marrow would be analogous to the double-positive thymocyte compartment in the thymus, and immunoglobulin light-chain gene recombination would correspond to the TCR V α rearrangements that continue after TCR surface deposition until positive selection¹⁹.

If there is continued light-chain gene recombination in IgM^{low} B cells in bone marrow and spleen due to ongoing RAG expression, how is immunoglobulin light-chain allelic exclusion established and editing regulated? The mechanisms regulating immunoglobulin light-chain allelic exclusion are not well defined, but continued RAG expression in immature B cells is consistent with a stochastic model²⁰. Although initial experiments with transgenic mice did not support this idea, indicating instead that a feedback mechanism might terminate immunoglobulin light-chain gene rearrangements, later work showed variable degrees of light-chain allelic exclusion depending in part on the specificity of the transgenic antibody (reviewed in ref. 21). In the extreme case of self-reactive antibody transgenes, there was increased RAG expression and receptor editing but no allelic exclusion^{1,2}. However, highly efficient editing was also found in B cells whose receptors bind self antigen with extremely low affinity and fail to induce high levels of RAG expression^{16,22}.

These apparently contradictory findings might be explained by the observation that IgM^{low} B cells in spleen and bone marrow express RAGs, and that the level of expression varies with the stage of development. We propose that the engagement of antigen receptors by self antigen during B-cell development might simply modulate the rate of transition between different compartments in the B-cell pathway^{23,24}. In the case of a non-self-reactive antibody transgene, B-cell development is accelerated²⁵ to the IgM^{high} stage and the level of RAG expression is decreased²⁶. In contrast, strong receptor crosslinking by self antigen in the bone marrow arrests cells in the small B220⁺IgM^{low}IgD⁺RAG^{high} stage, and may thereby raise the apparent level of RAG expression and the efficiency of editing^{23,24}. Finally, low-affinity receptor crosslinking may stimulate editing without major alterations in RAG levels by slowing the transition from the RAG-positive immature B-cell compartment to the RAG-negative long-lived B-cell compartment in the spleen^{10,16}. Although highly speculative, a model in which the B-cell receptor regulates the rates of cellular transitions late in B-cell development would be entirely consistent with previously documented alterations in B-cell development in mice that carry anti-self-reactive antibodies^{23,24}. A similar function for the antigen receptor has been documented in earlier stages of B-cell development^{25,27}. □

Methods

BACs and production of transgenic mice. A mouse BAC library (Research Genetics) was screened with a ³²P-labelled probe corresponding to base pairs

1138–2501 of mouse RAG1 (ref. 28). BACs were mapped by restriction digestion. GFP complementary DNA was inserted at the start codon of RAG2 by homologous recombination⁴. The shuttle vector for insertion of GFP into the RAG2 locus was created as follows: DNA fragments from –1191 to 0 and from 0 to 1016 (0 corresponds to the RAG2 start codon) were amplified separately by PCR from a BAC template²⁹. EGFP (Clontech) was amplified from plasmid DNA. PCR primers were designed to allow the three amplification products to anneal in the desired sequence and orientation. After a second round of PCR using only the two most distal primers, the product was cloned into the *Sal*I site of the shuttle vector⁴. A Kozak consensus sequence was introduced to improve translation, and there is a termination codon at the end of the GFP. The PCR primer pairs used to generate the shuttle vector insert were: 5′sal/r2, 5′-AGCAGACGAGTCGACACATCCACAAGCAGGAA GTACAC-3′; 3′r2/gfp, 5′-CCTCGCCCTTGCTCACCATGGTTGATTGTGAAT AGGTCTTTATCTGAA-3′; 5′r2/gfp, 5′-TTCAGATAAAGACCTATTACACA ATCAACCATGGTGAGCAAGGGCGAGG-3′; 3′gfp/r2, 5′-CCCCTGTTACC ATCTGCAGGGATTACTTGTACAGCTCGTCCATGCC-3′; 5′gfp/r2, 5′-GGC ATGGACGAGCTGTACAAGTAATCCTGCAGATGGTAACAGTGGG-3′; 3′r2/sal, 5′-GACATCAGAGTCGACATAGCCTGCTTATTGTCTCTCTG-3′.

Flow cytometry and immunohistochemistry. The following anti-mouse antibodies were used: biotin, phycoerythrin or allophycocyanin anti-B220 (RA3-6B2); biotin anti-CD43 (S7); phycoerythrin anti-HSA (M1/69); biotin anti-IgM (R6-60.2); biotin anti-Ig λ ₁₂ (R26-46); biotin anti-Ig κ (R5-240); GL7; and FcBlock (2.4G2) (all from Pharmingen); biotin 493 (ref. 7); phycoerythrin anti-mouse IgD (11–26; from Southern Biotechnology). Biotinylated antibodies were visualized with Streptavidin-RED670 (Gibco-BRL), Streptavidin-Per CP (Becton Dickinson) or Streptavidin-allophycocyanin (Pharmingen). GL7 was visualized with phycoerythrin mouse anti-rat IgM. In some experiments, propidium iodide was used to exclude dead cells.

Cell purification and culture. Bone marrow B cells from wild-type mice were enriched by positive selection using MACS B220 MicroBeads (Miltenyi Biotec) and stained with fluorescein isothiocyanate (FITC) anti-mouse IgM (Fab fragment, Jackson), PE anti-B220 and biotin anti-mouse CD43. B220^{low}CD43⁺IgM⁺, B220^{low}CD43⁺IgM^{low/med} and B220^{low}CD43⁺IgM^{high} fractions were separated on a FACS Vantage for RT-PCR (Becton Dickinson). Spleen B cells were enriched by negative selection with anti-mouse CD4 and anti-mouse CD8 coated magnetic beads (Dyna), then stained with FITC anti-mouse IgM (Fab fragment), PE anti-mouse HSA, biotin anti-mouse CD43 and APC anti-mouse B220. B220^{low}HSA^{high}CD43⁺ spleen immature B cells and B220^{high}HSA^{low}CD43⁺ mature B cells were sorted into IgM^{low}, IgM^{med} and IgM^{high} fractions. B220⁺GFP⁺ and B220⁺GFP⁺ B cells were sorted for western blot. B cells from NP-CGG immunized control and adoptively transferred Rag1^{−/−} mice were stained with anti-GL7 and anti-B220. GL7⁺B220⁺ and GL7⁺B220⁺ cells (6–7 × 10⁴) were directly sorted into TRIzol LS (Gibco-BRL) for RNA extraction. Purity of all of the sorted cells was 95–99% as determined by re-analysis. Spleen cells or purified B cells were cultured at 3 × 10⁶ cells ml^{−1} in IMDM (Gibco-BRL) supplemented with 10% fetal calf serum (Intergen), 50 μ M 2-mercaptoethanol, 2mM L-glutamine and LPS (25 μ g ml^{−1}), with or without IL4 (500 U ml^{−1}).

Adoptive transfers and immunization. For adoptive transfer, NG-BAC transgenic males (FVB/N) were bred with C57BL/6 (B6) females to obtain (NG-BAC × B6) F₁ mice. GFP⁺ or GFP[−] (NG-BAC × B6) F₁ B cells (2.5–6.5 × 10⁶) were purified and mixed with (NG-BAC × B6) F₁ T cells (1.5 × 10⁷) purified from transgene-negative littermates. Cell mixtures were injected intravenously into B6 Rag1^{−/−} recipients. T cells were primed by footpad injection of 30 μ g chicken IgG (Cappel) in incomplete Freund's adjuvant 1–2 weeks before use. Mice were immunized by intraperitoneal injection of 100 μ g NP-coupled chicken gamma globulin precipitated in alum.

Western blot. Purified cells were lysed as described³ and incubated overnight at 4°C with either polyclonal anti-mouse RAG2 or pre-immune rabbit serum (from P. Cortes). Western blots used polyclonal anti-mouse RAG2 and biotinylated anti-rabbit monoclonal IgG1 (Sigma, clone RG-16) developed with Western Blot Chemiluminescent Reagent Plus (NEN).

Enzyme-linked immunosorbent assay. NP-specific IgM and IgG₁ antibodies were detected using NP₁₆-BSA as the coating antigen. Horseradish peroxidase (HRP)-conjugated affinity-purified goat anti-mouse IgM (Jackson) and goat anti-mouse IgG₁-HRP (Southern Biotechnology) were used in conjunction

with HRP substrate kit (BIO-RAD). Results are expressed as absorbance at 405 nm of serum diluted 1:300.

RNA preparation, reverse transcription and PCR. RNA was extracted using TRIzol (GIBCO BRL) and reverse transcribed in 20 μ l with Superscript II (GIBCO BRL). RT-PCR reactions were performed with serial dilutions of cDNA template using AmpliTaq Gold (Perkin Elmer). The following PCR conditions and primers were used: Ig β , 30 s at 94 °C, 60 s at 60 °C, 45 s at 72 °C for 30 cycles, 5'-CCATGGCCACATGGTGTCTGTC-3' and 5'-CCTGAGTGGTTTGTGTAGCAGTG-3'; RAG1, 30 s at 94 °C, 60 s at 62 °C, 45 s at 72 °C, for 32 cycles, 5'-CAACCAAGTCGACACATTCTAGCACTC-3' and 5'-CACGTCGATCCGAAAATCCTGGCAATG-3'; endogenous RAG2, 30 s at 94 °C, 60 s at 60 °C, 45 s at 72 °C for 36 cycles, 5'-CACATCCACAAGCAGGAAGTACAC-3' and 5'-GGTTCAGGGACATCTCTACTAAG-3'; β -actin, 30 s at 94 °C, 60 s at 62 °C, 45 s at 72 °C for 25 cycles, 5'-TACCACTGGCATCGTGATGGACT-3' and 5'-TTTCTGCATCTGTGCGGCAAT. To detect RSS breaks, DNA preparation in agarose plugs, linker ligation, PCR and Southern blot hybridization were performed as described³⁰. Ig β PCR was performed using the following conditions and primers: 30 s at 94 °C, 30 s at 60 °C, 80 s at 72 °C for 32 cycles, 5'-TCACTGCTACACAACCAGTCAAG-3' and 5'-AGTTCGTGC-CACAGCTGTGG-3'. The Ig β amplification product was visualized using ethidium bromide.

Received 19 April; accepted 19 May 1999.

- Gay, D., Saunders, T., Camper, S. & Weigert, M. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J. Exp. Med.* **177**, 999–1008 (1993).
- Tiegs, S. L., Russell, D. M. & Nemazee, D. Receptor editing in self-reactive bone marrow B cells. *J. Exp. Med.* **177**, 1009–1020 (1993).
- Grawunder, U. *et al.* Down-regulation of RAG1 and RAG2 gene expression in preB cells after functional immunoglobulin heavy chain rearrangement. *Immunity* **3**, 601–608 (1995).
- Yang, X. W., Model, P. & Heintz, N. Homologous recombination based modification in *Escherichia coli* and germline transmission in transgenic mice of a bacterial artificial chromosome. *Nature Biotechnol.* **15**, 859–865 (1997).
- Lin, W. C. & Desiderio, S. Regulation of V(D)J recombination activator protein RAG-2 by phosphorylation. *Science* **260**, 953–959 (1993).
- Allman, D. M., Ferguson, S. E. & Canero, M. P. Peripheral B cell maturation. I. Immature peripheral B cells in adults are heat-stable antigenhi and exhibit unique signaling characteristics. *J. Immunol.* **149**, 2533–2540 (1992).
- Rollink, A. G., Andersson, J. & Melchers, F. Characterization of immature B cells by a novel monoclonal antibody, by turnover and by mitogen reactivity. *Eur. J. Immunol.* **28**, 3738–3748 (1998).
- Han, S., Zheng, B., Schatz, D. G., Spanopoulou, E. & Kelsoe, G. Neoteny in lymphocytes: Rag1 and Rag2 expression in germinal center B cells. *Science* **274**, 2094–7 (1996).
- Hikida, M. *et al.* Reexpression of RAG-1 and RAG-2 genes in activated mature mouse B cells. *Science* **274**, 2092–2094 (1996).
- Hertz, M., Kouskoff, V., Nakamura, T. & Nemazee, D. V(D)J recombinase induction in splenic B lymphocytes is inhibited by antigen-receptor signalling. *Nature* **394**, 292–295 (1998).
- Meffre, E. *et al.* Antigen receptor engagement turns off the V(D)J recombination machinery in human tonsil B cells. *J. Exp. Med.* **188**, 765–772 (1998).
- Forster, I. & Rajewsky, K. The bulk of the peripheral B-cell pool in mice is stable and not rapidly renewed from the bone marrow. *Proc. Natl Acad. Sci. USA* **87**, 4781–4784 (1990).
- Papavasiliou, F. *et al.* V(D)J recombination in mature B cells a new mechanism for diversification of antibody responses. *Science* **278**, 298–301 (1997).
- Russell, D. M. *et al.* Peripheral deletion of self-reactive B cells. *Nature* **354**, 308–311 (1991).
- Hartley, S. B. *et al.* Elimination from peripheral lymphoid tissues of self-reactive B lymphocytes recognizing membrane-bound antigens. *Nature* **353**, 765–769 (1991).
- Pelanda, R. *et al.* Receptor editing in a transgenic mouse model: site, efficiency, and role in B cell tolerance and antibody diversification. *Immunity* **7**, 765–775 (1997).
- Han, S. *et al.* V(D)J recombinase activity in a subset of germinal center B lymphocytes. *Science* **278**, 301–305 (1997).
- von Boehmer, H. & Fehling, H. J. Structure and function of the pre-T cell receptor. *Annu. Rev. Immunol.* **15**, 433–452 (1997).
- Turka, L. A. *et al.* Thymocyte expression of RAG-1 and RAG-2; termination by T cell receptor cross-linking. *Science* **253**, 778–781 (1991).
- Colclough, C., Perry, R. P., Karjalainen, K. & Weigert, M. Aberrant rearrangements contribute significantly to the allelic exclusion of immunoglobulin gene expression. *Nature* **290**, 372–378 (1981).
- Storb, U. in *Immunoglobulin Genes* (eds Honjo, T. & Alt, F.) 345–363 (Academic, San Diego, 1995).
- Melamed, D., Benschop, R. J., Cambier, J. C. & Nemazee, D. Developmental regulation of B lymphocyte immune tolerance compartmentalizes clonal selection from receptor selection. *Cell* **92**, 173–182 (1998).
- Hartley, S. B. *et al.* Elimination of self-reactive B lymphocytes proceeds in two stages: arrested development and cell death. *Cell* **72**, 325–335 (1993).
- Chen, C. *et al.* The site and stage of anti-DNA B-cell deletion. *Nature* **373**, 252–255 (1995).
- Nussenzweig, M. C. *et al.* A human immunoglobulin gene reduces the incidence of lymphomas in c-Myc-bearing transgenic mice. *Nature* **336**, 446–450 (1988).
- Costa, T. E. F., Suh, H. & Nussenzweig, M. C. Chromosomal position of rearranging gene segments influences allelic exclusion in transgenic mice. *Proc. Natl Acad. Sci. USA* **89**, 2205–2208 (1992).
- Kitamura, D., Roes, J., Kuhn, R. & Rajewsky, K. A B-cell deficient mouse by targeted disruption of the membrane exons of the immunoglobulin μ -chain gene. *Nature* **350**, 423–426 (1991).
- Schatz, D. G., Oettinger, M. A. & Baltimore, D. The V(D)J recombination activating gene, RAG-1. *Cell* **59**, 1035–1048 (1989).
- Oettinger, M. A., Schatz, D. G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* **248**, 1517–1523 (1990).
- Schlissel, M., Constantinescu, U. A., Morrow, T., Baxter, M. & Peng, A. Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. *Genes Dev.* **7**, 2520–2532 (1993).

Acknowledgements. We thank R. Monroe, C. Seidel and F. Alt for discussing unpublished data; P. Cortes for discussion and anti-RAG2 polyclonal antibodies; N. Heintz and W. Yang for help with BAC technology; and F. Isdell and M. Genova for cell sorting. W.Y. was supported by the NIH, the Surdna Foundation and the William Randolph Hearst Foundation. This work was funded in part by grants from the NIH to M.C.N.

Correspondence and requests for materials should be addressed to M.C.N. (e-mail: nussen@rockvax.rockefeller.edu).

A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic pattern formation

Haitao Zhu*†, Peter Kavsak†‡§, Shirin Abdollah‡||, Jeffrey L. Wrana‡§|| & Gerald H. Thomsen*

*Department of Biochemistry and Cell Biology and Institute for Cell and Developmental Biology, State University of New York, Stony Brook, New York 11794-5215, USA

‡Program in Developmental Biology, The Hospital for Sick Children, and §Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Canada M5G 1X8

† These authors contributed equally to this work.

The TGF- β superfamily of proteins regulates many different biological processes, including cell growth, differentiation and embryonic pattern formation^{1–3}. TGF- β -like factors signal across cell membranes through complexes of transmembrane receptors known as type I and type II serine/threonine-kinase receptors, which in turn activate the SMAD signalling pathway^{4,5}. On the inside of the cell membrane, a receptor-regulated class of SMADs are phosphorylated by the type-I-receptor kinase. In this way, receptors for different factors are able to pass on specific signals along the pathway: for example, receptors for bone morphogenetic protein (BMP) target SMADs 1, 5 and 8, whereas receptors for activin and TGF- β target SMADs 2 and 3. Phosphorylation of receptor-regulated SMADs induces their association with Smad4, the 'common-partner' SMAD, and stimulates accumulation of this complex in the nucleus, where it regulates transcriptional responses. Here we describe Smurf1, a new member of the Hect family of E3 ubiquitin ligases. Smurf1 selectively interacts with receptor-regulated SMADs specific for the BMP pathway in order to trigger their ubiquitination and degradation, and hence their inactivation. In the amphibian *Xenopus laevis*, Smurf1 messenger RNA is localized to the animal pole of the egg; in *Xenopus* embryos, ectopic Smurf1 inhibits the transmission of BMP signals and thereby affects pattern formation. Smurf1 also enhances cellular responsiveness to the Smad2 (activin/TGF- β) pathway. Thus, targeted ubiquitination of SMADs may serve to control both embryonic development and a wide variety of cellular responses to TGF- β signals.

To identify new components of the SMAD pathway, we performed a yeast two-hybrid screen⁶ using *Xenopus* Smad1. This yielded a protein that interacts with Smad1 and has significant homology to the Hect subclass of E3 ubiquitin ligases⁷, which selectively target proteins for ubiquitination and degradation by the 26S proteasome⁸. We named this gene and a closely related human homologue, *XSmurf1* and *hSmurf1*, respectively (for *Xenopus* and human SMAD ubiquitination regulatory factor-1). Both encoded proteins comprise 731 amino acids, share 91%

|| Present address: Program in Molecular Biology and Cancer, Samuel Lunenfeld Research Institute, Mt Sinai Hospital, 600 University Avenue, Toronto, Canada M5G 1X5.

sequence identity and are most closely related to Pub1, a ubiquitin ligase in *Schizosaccharomyces pombe* that regulates mitosis by targeting Cdc25 for proteasomal degradation⁹. Other members of this class of E3 ubiquitin ligases⁷ include Nedd4 and yeast Rsp5^{10,11}. Smurf1 shares distinctive structural features with these E3 ubiquitin ligases (Fig. 1a). These include an amino-terminal phospholipid/calcium-binding C2 domain^{12,13}, two WW domains¹⁴, which facilitate protein–protein interactions by binding to PPXY motifs on partner proteins, and a carboxy-terminal HECT domain⁷, which catalyses ubiquitin ligation onto target proteins.

In *Xenopus* embryonic development, *XSmurf1* mRNA was expressed from the egg stage to the swimming tadpole, with maximum levels observed in the stages from egg to gastrula (Fig. 1b). The bulk of *XSmurf1* mRNA in early development represents maternal transcripts accumulated during oogenesis. Whole-mount *in situ* hybridization (Fig. 1c) revealed that the maternal *XSmurf1* transcripts were localized to the upper, animal-

pole half of the egg and cleavage-stage blastula, which was confirmed by northern blot analysis on total animal and vegetal RNA from the blastula (data not shown). At gastrulation, *XSmurf1* mRNA was distributed uniformly in embryonic ectoderm and involuting mesoderm, but expression gradually localized to the nervous system (data not shown). At early tadpole stages, *XSmurf1* was expressed in the central nervous system, eye, branchial arches, kidney and somites (Fig. 1c). This embryonic expression pattern of *XSmurf1* coincides partly with the expression patterns of genes encoding Smad1 and BMP at similar stages^{15–18}.

Smurf1 contains a putative E3 ligase, or Hect domain, and is a candidate protein that interacts with Smad1. We therefore investigated whether Smurf1 regulates steady-state levels of Smad1 protein. In two mammalian cell lines, 293T and COS-1, co-expression of Smad1 with hSmurf1 produced a significant, dose-dependent decrease in steady-state levels of Smad1 protein. This was specific for hSmurf1, as overexpression of Nedd4, a related E3 ubiquitin ligase, had only a slight effect on Smad1 protein (Fig. 2b). Transient hSmurf1 expression also strongly reduced the amount of endogenous Smad1 in 293T cells (Fig. 2c), both in the absence and presence of BMP stimulation. Furthermore, co-expressing Smad1 together with the activated BMP type-I receptor ALK6 (also known as BMPRII) did not alter hSmurf1-dependent decreases in Smad1 levels (Fig. 2d). In the course of these latter experiments, we often observed decreased steady-state levels of the activated receptor in the presence of hSmurf1. This effect was not observed for activated TGF- β receptors (data not shown) and may be due to Smad1 recruiting some hSmurf1 to the activated receptor, with subsequent enhancement of receptor turnover. Together, these results show that expression of hSmurf1 causes dose-dependent decreases in Smad1 steady-state levels that occur independently of activation of Smad1 by the type-I BMP receptor.

To determine whether Smurf1 effects were exclusive to Smad1, we studied Smad2, a receptor-regulated SMAD that functions in TGF- β - and activin-signalling pathways. Relative to Smad1, hSmurf1 had a smaller effect on steady-state levels of Smad2 protein, which occurred only at the highest levels of hSmurf1 expression (Fig. 2e). Similarly, hSmurf1 had little or no effect on Smad3 or Smad4, but elicited a strong decrease in Smad5 protein, which is closely related to Smad1 (Fig. 2f). Thus, hSmurf1 regulates the steady-state levels of Smad1 and Smad5, two receptor-regulated SMADs that function in BMP signalling.

To investigate whether hSmurf1 regulates SMAD degradation, we analysed Smad1 turnover by pulse-chase experiments. In the absence of hSmurf1, Smad1 had a half-life of approximately 6 h. However, in the presence of hSmurf1, Smad1 turnover was enhanced, and displayed a half-life of less than 2 h (Fig. 3a). To determine whether decreases in Smad1 protein were dependent on the proteasome, we tested two inhibitors. In COS-1 cells, LLnL treatment prevented hSmurf1-dependent decreases in Smad1 (Fig. 3b). We were unable to assess LLnL in 293T cells owing to toxic effects, but the inhibitor lactacystin¹⁹ also suppressed hSmurf1-dependent loss of Smad1 protein in 293T cells (Fig. 3b). These data demonstrate that hSmurf1 induces Smad1 degradation through the proteasome.

We next tested whether hSmurf1 enhances Smad1 turnover through ubiquitination. We transfected 293T cells with HA-tagged ubiquitin²⁰ together with Flag-tagged Smad1, in the presence or absence of Myc-tagged hSmurf1. In the absence of hSmurf1, Smad1 displayed little or no detectable ubiquitination; however, upon co-transfection with hSmurf1 we observed the appearance of a ladder of ubiquitin-conjugated Smad1 products (Fig. 3c). To confirm that ubiquitination of Smad1 required the catalytic activity of the Hect domain in hSmurf1, we constructed a point mutant of hSmurf1 (hSmurf1(C710A)) that targets the cysteine residue that is thought to form a thioester bond with ubiquitin⁷. In contrast to wild-type hSmurf1, expression of hSmurf1 (C710A) did not yield

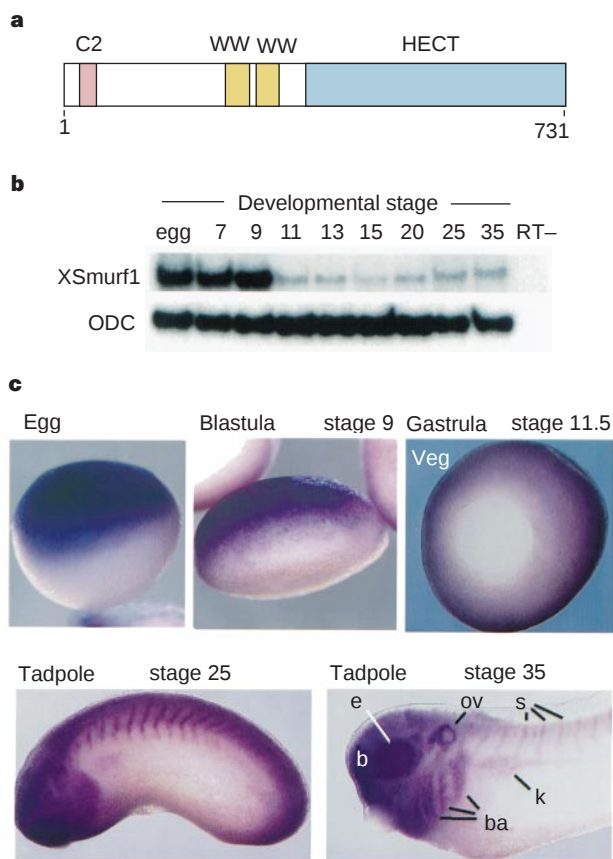


Figure 1 *Smurf1* encodes an E3 ubiquitin ligase expressed in early *Xenopus* development. **a**, Schematic of conserved protein domains found in *Xenopus* and human Smurf1 proteins: a lipid/Ca²⁺-binding (C2) domain at residues 22–37, WW protein interaction domains at residues 236–271 and 282–311, and a catalytic Hect domain at residues 347–731. See Supplementary Information for a detailed comparison of these proteins. **b**, Expression of *XSmurf1* in *Xenopus* embryonic development, assayed by RT-PCR. Embryonic stages³⁰ correspond to blastula (stages 7 and 9), gastrula (stages 11 and 13), neurula (stages 15 and 20) and tadpole (stages 25 and 35). Ornithine decarboxylase (ODC) expression normalizes RNA recovery and reaction integrity. RT- represents PCR on mock cDNA (no reverse transcriptase). **c**, Whole-mount *in situ* hybridization of *XSmurf1* in developing albino *Xenopus* embryos. In the egg and blastula, *XSmurf1* transcripts are localized to the animal pole (dark purple stain). Gastrula view is from the vegetal pole (Veg); yolk plug not stained. Lower panels, lateral views of stage 25 and 35 tadpoles; brain (b), eye (e), otic vesicle (ov), somites (s), branchial arches (ba) and kidney (k).

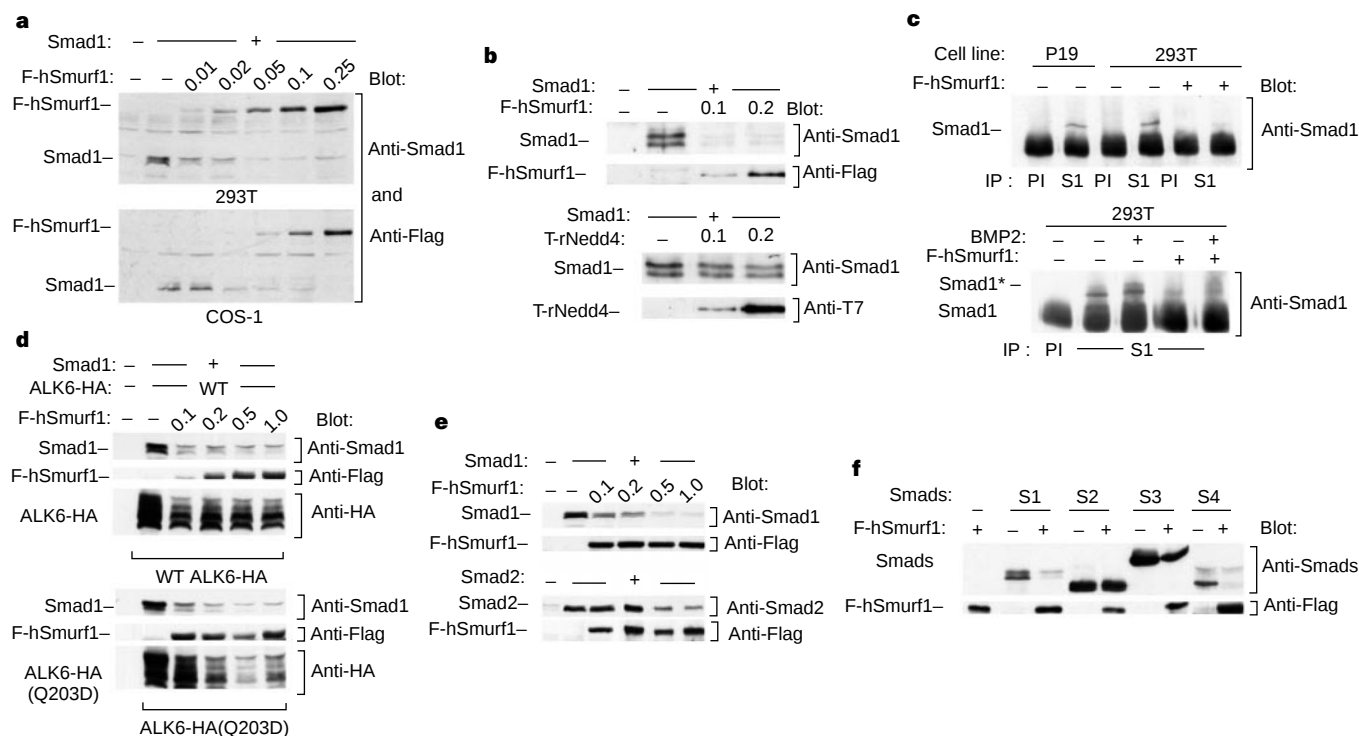


Figure 2 Smurf1 expression decreases the steady-state level of Smad1 and Smad5. COS-1 (**a**) or 293T (**a-f**) cells were transiently transfected with the indicated mammalian expression constructs. In **a**, **b**, **d** and **e**, the amount of Flag-tagged hSmurf1 (F-hSmurf1) DNA is shown in micrograms. Steady-state protein levels were determined by immunoblotting aliquots of total cell lysates using the appropriate antibodies as shown. **a**, **b**, hSmurf1, but not Nedd4, decreases Smad1 steady-state levels. Smad1 protein levels were determined in cells transfected with Smad1 and increasing concentrations of Flag-tagged hSmurf1 (**a**) or T7-tagged rat Nedd4 (T-rNedd4)¹³(**b**). **c**, hSmurf1 decreases endogenous Smad1 levels. Lysates prepared from P19, 293T (-) or 293T cells transiently expressing Flag-tagged hSmurf1 (+) were subjected to immunoprecipitation (IP) with preimmune (PI) or anti-Smad1 (S1) antibodies, followed by anti-Smad1

immunoblotting. Bottom panel: 293T cells were incubated in the presence (+) or absence (-) of BMP2 for 30 min before analysis of Smad1 levels. The migration of phosphorylated Smad1 is indicated (Smad1*). F-hSmurf1 expression was confirmed by immunoblotting total cell lysates (not shown). **d**, hSmurf1 reduces Smad1 steady-state levels independent of receptor activation. Smad1 levels were assessed in cells co-expressing Smad1, F-hSmurf1 and wild-type (WT) or activated (Q203D) ALK6/BMPRII-HA receptors (anti-HA). **e**, **f**, hSmurf1 preferentially targets Smad1 and Smad5. Steady-state levels of transfected Smad1 or Smad2 protein were examined in lysates from cells expressing increasing amounts of F-hSmurf1 (**e**). The steady-state level of transfected Smad1, Smad3, Smad4 or Smad5 in the presence (+) or absence (-) of F-hSmurf1 was assessed by immunoblotting cell lysates with the appropriate antibody (anti-Smads).

ubiquitinated Smad1 (Fig. 3d) and, moreover, hSmurf1 (C710A) did not induce decreased steady-state levels of Smad1 protein compared with wild-type hSmurf1 (Fig. 3e). Thus, hSmurf1 controls Smad1 steady-state levels by inducing poly-ubiquitination of Smad1 through its Hect domain, with subsequent degradation through the proteasome.

To assess how Smurf1 selectively targets Smad1 and Smad5 for degradation, we investigated the interaction of Smurf1 with SMAD proteins. Assays performed *in vitro* showed that XSmurf1 selectively bound to Smad1, but not to Smad2 and Smad4 (Fig. 4a). In mammalian cells, we were unable to detect interactions between wild-type hSmurf1 and Smad1. However, the interaction between the Smurf1 ubiquitin ligase and its Smad1 substrate may be transient in intact cells. Therefore we examined the ubiquitin-ligase mutant of hSmurf1 and found that hSmurf1(C710A) formed a stable complex with Smad1 and Smad5, but associated only weakly with Smad2 (Fig. 4b). Furthermore, this interaction was unaffected by co-expression of the constitutively active BMP type-I receptor, ALK2 (ref. 21, and data not shown), which is consistent with the notion that hSmurf1 regulates Smad1 turnover independent of BMP signalling. The linker regions of receptor-regulated SMADs contain a PPXY sequence, which is a conserved motif recognized by WW domains¹⁴, such as those found in both XSmurf1 and hSmurf1. Consequently, we tested whether hSmurf1 would interact with Smad1 in which the PY motif was deleted (residues 223–227,

inclusive). Unlike wild-type Smad1, there was little if any association between Smad1(ΔPY) and hSmurf1(C710A) (Fig. 4c), and Smad1(ΔPY) was resistant to hSmurf1-mediated degradation (Fig. 4d). It therefore seems that Smurf1 associates specifically with Smad1 and Smad5 through interactions between the PY motif in Smad1 and the WW domains of Smurf1.

In cultured cells, hSmurf1 downregulates the amount of BMP-pathway-specific Smad1 and Smad5 proteins. We therefore tested whether XSmurf1 might regulate BMP signalling in the ectodermal and mesodermal embryonic germ layers of the *Xenopus* embryo. In mesoderm of the ventral marginal zone (VMZ), BMP signals specify tissues such as blood and mesenchyme. However, if BMP signals are blocked in the VMZ, dorsal mesodermal tissues such as muscle develop in a process referred to as dorsalization³. Expression of ectopic XSmurf1 in the VMZ triggered formation of ectopic axial structures in the ventral side of resulting tadpoles (Fig. 5a), which is characteristic of dorsalization resulting from BMP inhibition. This correlated with inhibition of expression of markers for blood differentiation, and was concomitant with activation of muscle-differentiation markers in treated VMZ explants (Fig. 5a, right panel). In both intact tadpoles and VMZ explants, co-expression of ectopic Smad1 together with XSmurf1 reversed the dorsalizing effects of XSmurf1, demonstrating specificity for the BMP pathway. Furthermore, overexpression of XSmurf1 in cells of the dorsal marginal zone (DMZ), which forms head and axial mesodermal

tissues, had no effect (data not shown). Smad2 mediates activin/nodal/Vg1 signals during development of the *Xenopus* dorsal axis^{3,5,22}. Therefore this latter observation is consistent with our findings in cultured cells that hSmurf1 does not target Smad2 to any appreciable extent.

In the ectodermal germ layer of *Xenopus* embryos, endogenous BMP signals specify epidermis, but if BMP signals are reduced or eliminated, the ectoderm differentiates into cement gland or neural tissue, respectively²³. Furthermore, varying Smad1 or Smad5 levels in *Xenopus* ectoderm also alters cell fate^{18,24,25}. Thus, the localization of *XSmurf1* mRNA in the *Xenopus* animal pole indicates that XSmurf1 may regulate ectodermal differentiation and pattern by modulating BMP signalling. We therefore overexpressed XSmurf1 in animal pole tissue (animal caps) and found that XSmurf1 triggered neural and cement-gland differentiation, which is characteristic of inhibited BMP signalling. These effects were reversed by co-expressing Smad1 (Fig. 5b). We suggest that a normal embryonic function of XSmurf1 is to provide an intracellular, cell-autonomous mechanism for antagonizing BMP signalling through alterations in levels of Smad1 or Smad5 protein. This regulation might complement the function of BMP inhibitors secreted by the Spemann organizer, such as noggin, chordin and follistatin, which bind and block BMP ligands.

To test whether XSmurf1 specifically inhibits the biological effects of BMP signals at the SMAD level, we co-expressed XSmurf1 in

combination with Smad1 or Smad2 in *Xenopus* animal caps. Overexpression of particular receptor-regulated SMADs in animal caps mimics the effects of their corresponding TGF- β ligands or activated receptors. Thus, injection of *Smad1* mRNA induces ventral/posterior mesoderm, like BMP ligands, whereas injection of *Smad2* mRNA induces dorsal (Spemann organizer) mesoderm, like activin, Vg1 and nodal ligands. Injection of *Smad1* mRNA alone (1 ng) in animal caps induced the ventral/posterior mesoderm-specific genes *Xhox3* and *Xcad1*. However, co-injection of *XSmurf1* mRNA blocked Smad1-dependent induction of these genes at all *XSmurf1* doses tested (Fig. 6a), demonstrating that XSmurf1 antagonizes Smad1. We next tested XSmurf1 effects on Smad2 activity by injecting animal caps with a limiting dose of *Smad2* mRNA (50 pg) that was sufficient to induce muscle-specific *myoD*, a marker of dorsal/lateral mesoderm, but insufficient to induce *goosecoid*, a marker gene specific to the Spemann organizer (Fig. 6b,c). Co-expression of any dose of *XSmurf1* (up to 400 pg mRNA) did not interfere with Smad2-dependent induction of the *myoD* gene (Fig. 6b), which is consistent with our data showing that human or *Xenopus* Smurf1 does not efficiently target Smad2. When we monitored *goosecoid*, we observed that, as the dose of *XSmurf1* mRNA was increased from 50 to 400 pg, *goosecoid* gene expression was triggered (Fig. 6b). This mimicked the effects of injecting high doses of *Smad2* alone. Injection of 50 pg *Smad2* and 100 pg *XSmurf1* mRNA triggered *goosecoid* expression to almost the same degree as a

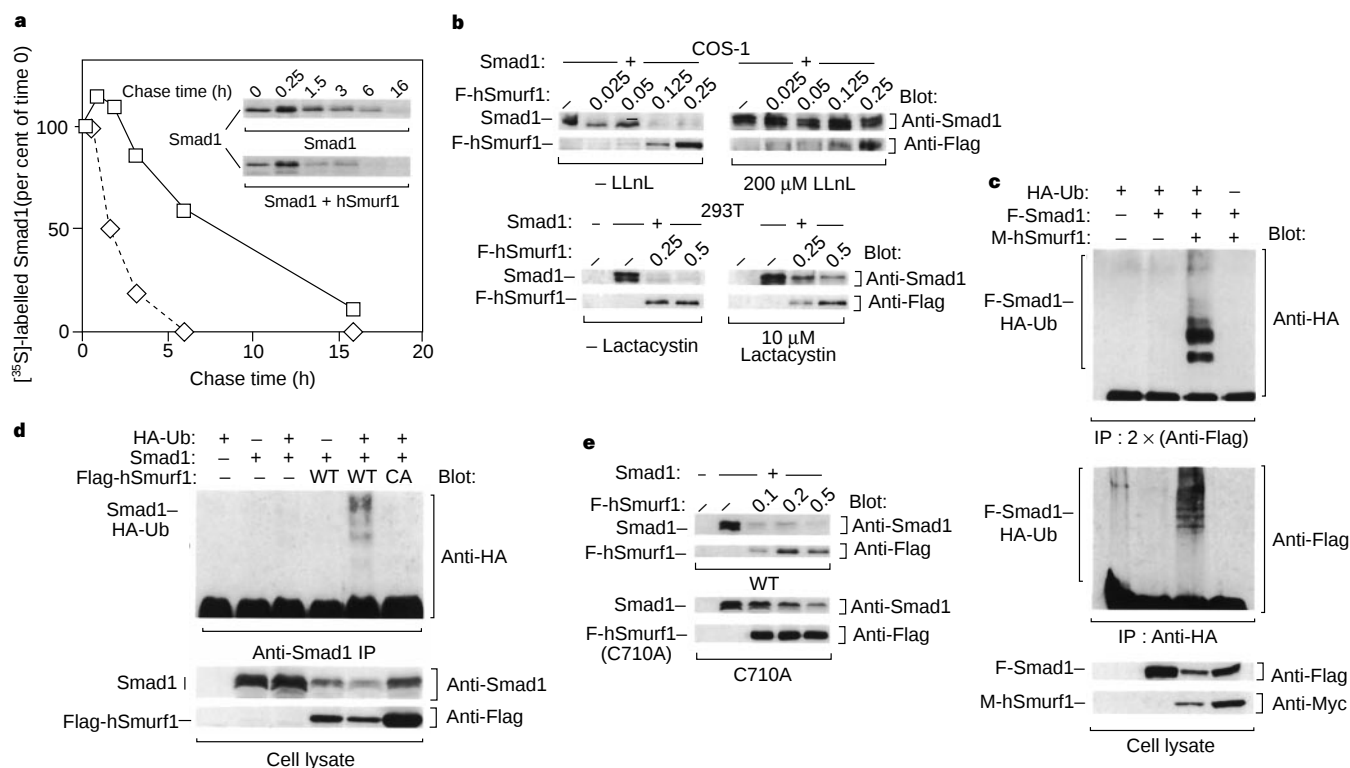


Figure 3 hSmurf1 induces Smad1 turnover and ubiquitination. **a**, COS-1 cells transfected with Smad1 and Flag-tagged hSmurf1 (F-hSmurf1) were pulse-labelled with ³⁵S-methionine and then chased for the indicated times in media containing unlabelled methionine. ³⁵S-labelled Smad1 in anti-Smad1 immunoprecipitates (inset) was quantified by phosphorimaging and the levels in control cells (squares) and hSmurf1-expressing cells (diamonds) are plotted relative to the amount present at time 0. **b**, Effect of proteasome inhibitors on Smad1 stability. The indicated cells, transfected with Smad1 and increasing amounts of F-hSmurf1, were treated overnight with or without 200 μ M LLnL (top panel) or 10 μ M lactacystin (bottom panel), before analysis of Smad1 steady-state levels. **c**, Ubiquitination of Smad1 in 293T cells. Lysates from cells transfected with HA-tagged ubiquitin (HA-Ub), Flag-tagged Smad1 (F-Smad1) and/or Myc-

tagged hSmurf1 (M-hSmurf1) were subjected to immunoprecipitation followed by immunoblotting as indicated. Ubiquitinated species of Smad1 are indicated (F-Smad1-HA-Ub). **d**, **e**, Ubiquitination and loss of Smad1 requires activity of the Smurf1 Hect domain. Lysates from cells transiently expressing HA-ubiquitin, Smad1 and either wild-type (WT) or the C710A mutant of F-hSmurf1 (CA) were subjected to anti-Smad1 immunoprecipitation (anti-Smad1 IP) followed by anti-HA immunoblotting (anti-HA blot) (**d**). Smad1 protein levels (anti-Smad1) were assessed in total cell lysates from 293T cells transfected with Smad1 and increasing amounts of wild-type (WT) or F-hSmurf1(C710A) (**e**). In **b**–**e**, protein expression was confirmed by immunoblotting aliquots of total cell lysates (bottom panels).

fivefold higher (250 pg) dose of *Smad2* alone (Fig. 6c). The induction of *gooseoid* in the presence of *XSmurf1* was *Smad2*-dependent, as *XSmurf1* alone did not activate *gooseoid* at any dose tested. Thus, *XSmurf1* enhanced the sensitivity of animal-cap cells to *Smad2*. Together, these data show that *Smurf1* alters the competence of animal-pole cells to respond to different TGF- β signalling pathways by simultaneously inhibiting BMP signals and enhancing activin-like signals. Because receptor-regulated SMADs in the BMP and activin pathways may compete for the common SMAD, *Smad4* in *Xenopus* embryos^{26,27}, *Smurf1* may sensitize animal pole cells to *Smad2*/activin by reducing endogenous *Smad1* or *Smad5*, thus liberating *Smad4* and promoting formation of *Smad2*–*Smad4* complexes.

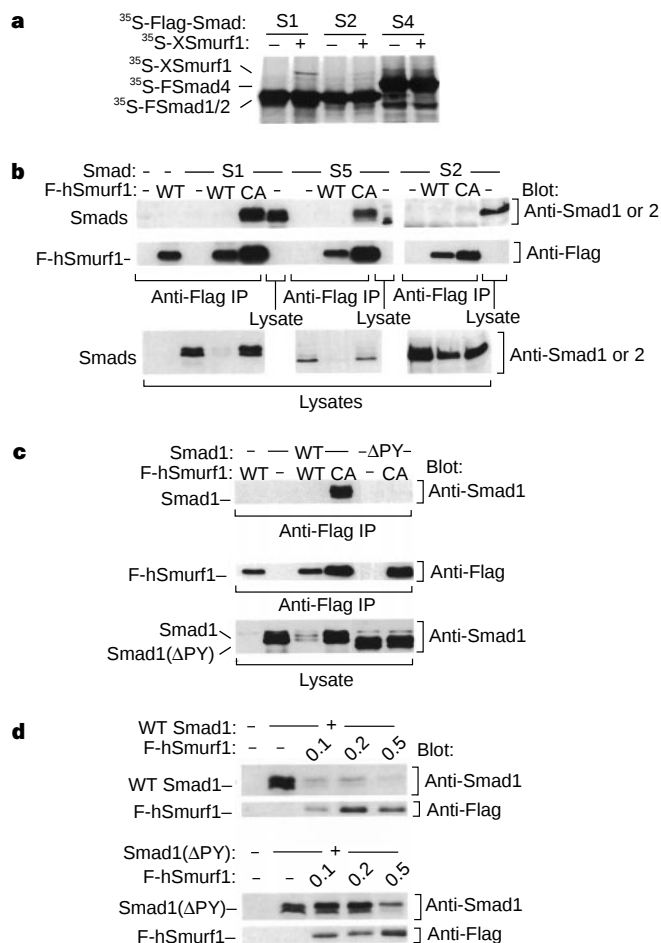


Figure 4 Interaction of Smurf1 and Smads. **a**, *XSmurf1* interacts with *Smad1* *in vitro*. Interaction between *XSmurf1* and *Smad1* was assessed by incubating *in vitro*-translated, [³⁵S]Met-labelled *XSmurf1* with *in vitro*-translated, [³⁵S]Met-labelled Flag-tagged *Smad1*, *Smad2* or *Smad4* immobilized on anti-Flag affinity-gel matrices. **b**, *hSmurf1* interacts selectively with *Smad1* and *Smad5* in mammalian cells. 293T cells transfected with *Smad1* (S1), *Smad5* (S5) or *Smad2* (S2) and either wild-type (WT) or the ubiquitin-ligase mutant C710A (CA) of Flag-tagged *hSmurf1* (F-*hSmurf1*) were subjected to anti-Flag immunoprecipitation followed by immunoblotting with anti-*Smad1/5* or anti-*Smad2* polyclonal antibody, as appropriate. The levels of F-*hSmurf1* in the immunoprecipitates and *Smad* levels in total cell lysates are shown as indicated. **c**, **d**, The PY motif in *Smad1* is an important determinant for interaction with *Smurf1*. Lysates from 293T cells transfected with *Smad1*, *Smad1*(Δ PY) and either wild-type (WT) or mutant (CA) F-*hSmurf1* were subjected to anti-Flag immunoprecipitation followed by immunoblotting with anti-*Smad1* antibodies to detect associated *Smad1* (**c**). Steady-state levels of transfected *Smad1* or *Smad1*(Δ PY) were assessed by immunoblotting total cell lysates from 293T cells expressing increasing amounts of F-*hSmurf1*.

In conclusion, we have shown that TGF- β signalling can be controlled by selective ubiquitination of SMAD signalling molecules. These findings have important implications for the regulation of cell fate in development and, more generally, cell responses to TGF- β signals in diverse biological contexts. The identification of *Smurf1*, a SMAD-specific E3 ubiquitin ligase, provides the first link between TGF- β signalling and the ubiquitination system. *Smurf1* blocks intracellular BMP signals by specifically targeting *Smad1* and *Smad5* for ubiquitination and proteasomal degradation, independent of BMP receptor activation. This indicates that *Smurf1* does not function downstream of activated SMADs to turn off BMP signals but rather controls cell competence to respond to BMPs. In particular, we demonstrate that *Smurf1* reduces *Smad1* and *Smad5* protein levels in cultured cells and affects embryonic patterning by BMP signals in the ectoderm and mesoderm of *Xenopus* embryos. Moreover, changes in the levels of *Smurf1* can alter the competence of embryonic cells to respond to different TGF- β signalling pathways, as increased amounts of *Smurf1* not only inhibit cellular responses to the *Smad1/5* (BMP) pathway, but also simultaneously enhance sensitivity to the *Smad2* (activin/TGF- β) pathway. The selective nature of *Smurf1*-mediated SMAD

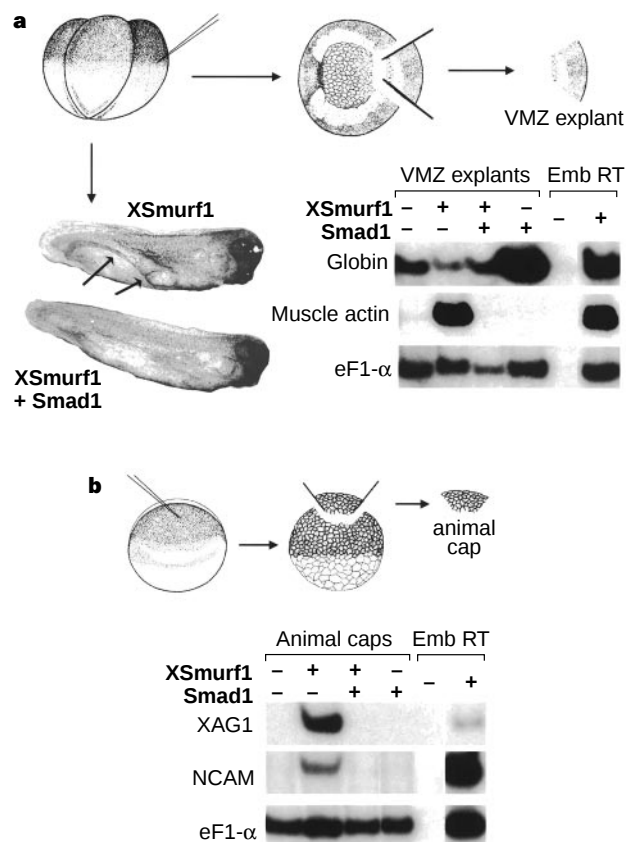


Figure 5 *Smurf1* dorsalizes ventral mesoderm and neuralizes ectoderm of *Xenopus* embryos. **a**, Two ventral blastomeres of four-cell *Xenopus* embryos were injected in the marginal zone with *XSmurf1* mRNA (50 pg per cell), with (+) or without (-) 100 pg *Smad1*. Left, tadpoles developed from injected embryos formed ectopic, dorsal axial structures (arrows). Right, in a parallel experiment, ventral marginal zone (VMZ) explants were excised at early gastrulation and scored by RT-PCR at tadpole stage 28 for expression of the indicated marker genes. **b**, Animal poles of fertilized eggs were injected with 100 pg of *XSmurf1*, with (+) or without (-) *Smad1* mRNA, or were uninjected. Animal-cap explants were analysed by RT-PCR at gastrula stage 11 for expression of *XAG*, a cement gland marker, and *NCAM*, a general neural marker. In both **a** and **b**, the two lanes on the far right are control reactions on total embryonic RNA, with (+) or without (-) reverse transcription.

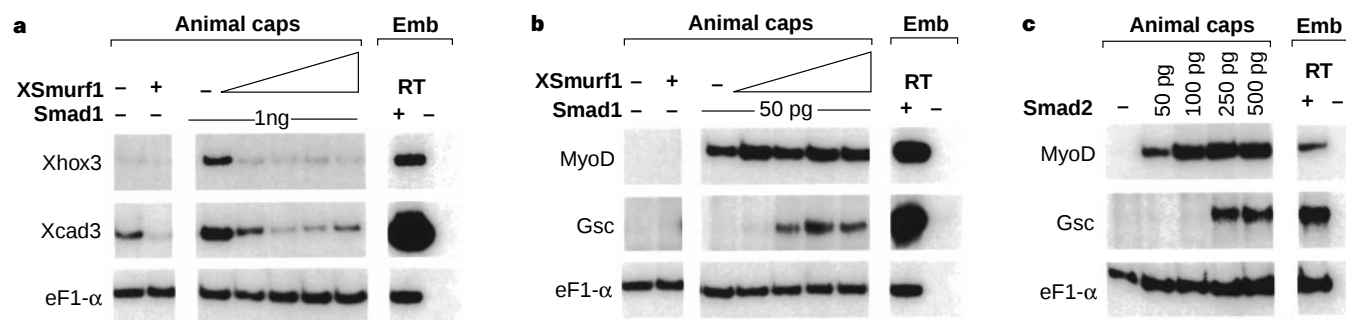


Figure 6 Smurf1 alters embryonic cell competence to respond to Smad1 and Smad2. **a**, XSmurf1 blocks ventral mesoderm induction by Smad1. Animal poles from fertilized eggs were injected with 1 ng Smad1 mRNA alone, or together with increasing amounts of XSmurf1 mRNA (gradient), or were not injected. Animal caps were excised from mid-blastula-stage embryos, cultured to late gastrula stage, then scored for Xhox3 and Xcad3 expression by RT-PCR. In lanes from left to right, injected doses of XSmurf1 mRNA were 0, 100, 0, 50, 100, 200 and 400 pg, respectively. **b**, XSmurf1 enhances dorsal mesoderm induction by Smad2. Animal

poles from fertilized eggs were injected with a constant amount (50 pg) of Smad2 mRNA either alone, or together with increasing amounts of XSmurf1 mRNA (as in **a**), cultured as above, then scored for gooseoid and myoD expression by RT-PCR. Note that gooseoid expression was triggered from undetectable levels as the dose of XSmurf1 was increased. **c**, Dose-response of animal caps to Smad2. Amounts of injected Smad2 mRNA and RT-PCR analysis were as in **b**. Controls in all panels were as in Fig. 5.

ubiquitination and the distinct effects of Smurf1 on cell responses to TGF- β -like factors provides a new mechanism for regulating cell growth, embryonic development, tissue stasis and other processes controlled by SMAD signalling pathways. □

Methods

Cloning of Smurf1 genes and construction of complementary DNA. A *Xenopus* Smad1 cDNA¹⁸ was cloned into the pGBT9 vector and used to screen a *Xenopus* oocyte cDNA library (Clontech) by the yeast two-hybrid method⁶. A partial cDNA we isolated was used to screen a *Xenopus* stage 9 (blastula) cDNA library to obtain a full-length XSmurf1 cDNA clone (GenBank accession number AF169310). A human Smurf1 cDNA encoding all but the first eight amino acids was identified in the expressed-sequence-tag database (AA292123), and full-length hSmurf1 was reconstituted using a fragment of XSmurf1 cDNA encoding the first eight amino acids. hSmurf1 cDNA was Flag-tagged at its N terminus. The polymerase chain reaction (PCR) was used to generate a ubiquitin-ligase mutant of hSmurf1 in which cysteine 710 was replaced with alanine and the Δ PPY mutant of Smad1 in which the PPPAY sequence between amino-acid residues 223–227 was deleted.

Immunoprecipitations and immunoblotting. For immunoprecipitation assays, *Xenopus* Smad1 (ref. 18), mouse Smad4 and human Smad2 (ref. 28) were Flag-tagged at their N termini and translated *in vitro* (rabbit reticulocyte extracts; Promega) in the presence of ³⁵S-methionine. The Flag-tagged SMADs were bound to anti-Flag antibody-conjugated beads (Kodak), washed in co-immunoprecipitation buffer (10 mM Tris, pH 7.5, 90 mM NaCl, 1 mM EDTA, 1% Triton-X 100, 10% glycerol, 1 mM phenylmethylsulphonylfluoride) then incubated with [³⁵S]Met-labelled hSmurf1 in the same buffer. After washing in co-immunoprecipitation buffer, proteins were eluted in gel-loading buffer, separated by SDS-polyacrylamide gel electrophoresis (PAGE) and visualized by autoradiography.

For studies in mammalian cells, COS-1 and 293T cells were transiently transfected using lipofectAMINE (Gibco/BRL) or calcium phosphate precipitation, respectively. Immunoprecipitations and immunoblotting were done using anti-HA (12CA5, Boehringer), anti-Myc (9E10), anti-Flag M2 (Sigma) or anti-T7 (Novagen) monoclonal antibodies, or anti-Smad1/5 or anti-Smad2/3 polyclonal antibodies, as described previously²¹. Detection was achieved using the appropriate horseradish peroxidase (HRP)-conjugated goat anti-mouse or goat anti-rabbit secondary antibodies and enhanced chemiluminescence (Amersham). In experiments that used proteasome inhibitors, transfected COS1 or 293T cells were incubated overnight with 200 μ M LLN1 or 10 μ M lactacystin, respectively, as described previously²⁹, before analysis of Smad1 protein levels.

Embryo methods and RT-PCR. All *Xenopus* embryo methods and PCR coupled with reverse transcription (RT-PCR) were done as described¹⁸. Primers for RT-PCR on XSmurf1 were 5'-GTCCTGTGACTGGAACCC-3' (sense) and

5'-GAGGACTGCTAGACAAT-3' (antisense), with 5' ends respectively located at positions 482 and 726 in the cDNA. Other primers used are described on the *Xenopus* Molecular Marker Homepage (<http://vize222.zo.utexas.edu>).

Pulse-chase analysis. COS-1 cells were transfected as indicated, and two days post-transfection the cells were labelled for 10 min at 37 °C with 50 μ Ci ³⁵S-methionine per ml in methionine-free Dulbecco's modified Eagle's medium (DMEM; Pulse). Cell layers were then washed two times and incubated in DMEM + 10% fetal calf serum (FCS) for the indicated time periods (Chase). At each time point of the chase, cell lysates were immunoprecipitated with an anti-Smad1 polyclonal antibody, resolved by SDS-PAGE and visualized by autoradiography. A Phosphorimager (Molecular Dynamics) was used to quantify metabolically labelled Smad1 present at each time point.

Ubiquitination assay. 293T cells were transfected with HA-tagged ubiquitin, Flag-tagged or untagged Smad1, and either Flag-tagged or Myc-tagged hSmurf1 or Flag-hSmurf1 (C710A) as indicated. Two days post-transfection, cell lysates were subjected to either anti-Flag or anti-HA immunoprecipitation. Anti-Flag immunoprecipitates were boiled in 1% SDS for 5 min, diluted with TNTe (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.1% TritonX-100) and subjected to a second anti-Flag immunoprecipitation before immunoblotting with anti-HA to detect HA-ubiquitin conjugated to Smad1. Anti-HA immunoprecipitates were immunoblotted with anti-Flag to visualize HA-ubiquitinated Smad1. Expression levels of transiently expressed Smad1 and hSmurf1 were analysed by immunoblotting aliquots of total cell lysates with the appropriate antibodies.

Received 12 April; accepted 3 June 1999.

- Kingsley, D. M. The TGF- β superfamily: new members, new receptors, and new genetic tests of function in different organisms. *Genes Dev.* **8**, 133–146 (1994).
- Moses, H. L. & Serra, R. Regulation of differentiation by TGF- β . *Curr. Opin. Genet. Dev.* **6**, 581–586 (1996).
- Harland, R. & Gerhart, J. Formation and Function of Spemann's organizer. *Annu. Rev. Cell Biol.* **13**, 611–667 (1997).
- Massague, J. TGF- β signal transduction. *Annu. Rev. Biochem.* **67**, 753–791 (1998).
- Whitman, M. Smads and early developmental signaling by the TGF β superfamily. *Genes Dev.* **12**, 2445–2462 (1998).
- Bartel, P. & Fields, S. Analyzing protein-protein interactions using two-hybrid system. *Methods Enzymol.* **254**, 241–263 (1995).
- Huibregtse, J. M., Scheffner, M., Beaudenon, S. & Howley, P. M. A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase. *Proc. Natl Acad. Sci. USA* **92**, 2563–2567 (1995). [Published erratum appears in *Proc. Natl Acad. Sci. USA* **92**, 5249 (1995).]
- Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Nefsky, B. & Beach, D. Pub1 acts as an E6-AP-like protein ubiquitin ligase in the degradation of cdc25. *EMBO J.* **15**, 1301–1312 (1996).
- Scheffner, M., Huibregtse, J. M., Vierstra, R. D. & Howley, P. M. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* **75**, 495–505 (1993).
- Hein, C., Springael, J., Volland, C., Haguenaer-Tsapir, R. & Andre, B. NPI1, an essential yeast gene involved in induced degradation of Gap1 and Fur4 permeases, encodes the Rsp5 ubiquitin-protein ligase. *Mol. Microbiol.* **18**, 77–87 (1995).
- Nalefski, E. A. & Falke, J. J. The C2 domain calcium-binding motif: structural and functional diversity. *Protein Sci.* **5**, 2375–2390 (1996).
- Plant, P. J., Yeager, H., Staub, O., Howard, P. & Rotin, D. The C2 domain of the ubiquitin protein ligase Nedd4 mediates Ca²⁺-dependent plasma membrane localization. *J. Biol. Chem.* **272**, 32329–32336 (1997).

14. Rotin, D. WW (WWP) domains: from structure to function. *Curr. Top. Microbiol. Immunol.* **228**, 115–133 (1998).
15. Fainsod, A., Steinbeisser, H. & De Robertis, E. M. On the function of BMP-4 in patterning the marginal zone of the *Xenopus* embryo. *EMBO J.* **13**, 5015–5025 (1994).
16. Hemmati-Brivanlou, A. & Thomsen, G. H. Ventral mesodermal patterning in *Xenopus* embryos: expression patterns and activities of BMP-2 and BMP-4. *Dev. Genet.* **17**, 78–89 (1995).
17. Nishimatsu, S. & Thomsen, G. H. Ventral mesoderm induction and patterning by BMP heterodimers in *Xenopus* embryos. *Mech. Dev.* **74**, 75–88 (1997).
18. Thomsen, G. H. *Xenopus* mothers against decapentaplegic is an embryonic ventralizing agent that acts downstream of the BMP-2/4 receptor. *Development* **122**, 2359–2366 (1996).
19. Fenteany, G. *et al.* Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin. *Science* **268**, 726–731 (1995).
20. Treier, M., Staszewski, L. M. & Bohmann, D. Ubiquitin-dependent c-Jun degradation in vivo is mediated by the delta domain. *Cell* **78**, 787–798 (1994).
21. Macias-Silva, M., Hoodless, P. A., Tang, S. J., Buchwald, M. & Wrana, J. L. Specific activation of Smad1 signaling pathways by the BMP7 type I receptor, ALK2. *J. Biol. Chem.* **273**, 25628–25636 (1998).
22. Hoodless, P. A. *et al.* Dominant-negative Smad2 mutants inhibit activin/Vg1 signaling and disrupt axis formation in *Xenopus*. *Dev. Biol.* **207**, 364–379 (1999).
23. Hemmati-Brivanlou, A. & Melton, D. Vertebrate embryonic cells will become nerve cells unless told otherwise. *Cell* **88**, 13–17 (1997).
24. Wilson, P. A., Lagna, G., Suzuki, A. & Hemmati-Brivanlou, A. Concentration-dependent patterning of the *Xenopus* ectoderm by BMP4 and its signal transducer Smad1. *Development* **124**, 3177–3184 (1997).
25. Suzuki, A., Chang, C., Yingling, J. M., Wang, X. F. & Hemmati-Brivanlou, A. Smad5 induces ventral fates in *Xenopus* embryos. *Dev. Biol.* **184**, 402–405 (1997).
26. Lagna, G., Hata, A., Hemmati-Brivanlou, A. & Massague, J. Partnership between DPC4 and SMAD proteins in TGF- β signalling pathways. *Nature* **383**, 832–836 (1996).
27. Candia, A. F. *et al.* Cellular interpretation of multiple TGF- β signals: intracellular antagonism between activin/BVg1 and BMP-2/4 signaling mediated by Smads. *Development* **124**, 4467–4480 (1997).
28. Eppert, K. *et al.* MADR2 maps to 18q21 and encodes a TGF β regulated MAD-related protein that is functionally mutated in colorectal carcinoma. *Cell* **86**, 543–552 (1996).
29. Ward, C. L., Omura, S. & Kopito, R. R. Degradation of CFTR by the ubiquitin-proteasome pathway. *Cell* **83**, 121–127 (1995).
30. Nieuwkoop, P. D. & Faber, J. *Normal Table of Xenopus laevis (Daudin)* (North Holland, Amsterdam, 1967).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London Editorial office of Nature.

Acknowledgements. We thank P. Gergen, G. Golling, N. Hollingsworth and B. Li for reagents and advice on the yeast two-hybrid system; D. Rotin and D. Bohmann for reagents; D. Rotin for helpful discussions; and our colleagues who provided useful comments on the manuscript. This work was supported by the MRC and the NCIC with funds from the Terry Fox run (to J.L.W.), and the NIH, NSF and a Cancer Classic Award from the University Medical School (to G.H.T.). P.K. is a recipient of an MRC Studentship and J.L.W. is an MRC Scholar.

Correspondence and requests for materials should be addressed to G.H.T. (e-mail: ghthomsen@notes.cc.sunysb.edu).

Trigger factor and DnaK cooperate in folding of newly synthesized proteins

Elke Deuerling, Agnes Schulze-Specking, Toshifumi Tomoyasu, Axel Mogk & Bernd Bukau

Institut für Biochemie und Molekularbiologie, Hermann-Herder-Strasse 7, 79104 D-Freiburg, Germany

The role of molecular chaperones in assisting the folding of newly synthesized proteins in the cytosol is poorly understood. In *Escherichia coli*, GroEL assists folding of only a minority of proteins¹ and the Hsp70 homologue DnaK is not essential for protein folding or cell viability at intermediate growth temperatures². The major protein associated with nascent polypeptides is ribosome-bound trigger factor^{3,4}, which displays chaperone and prolyl isomerase activities *in vitro*^{3,5,6}. Here we show that Δ tig::kan mutants lacking trigger factor have no defects in growth or protein folding. However, combined Δ tig::kan and Δ dnaK mutations cause synthetic lethality. Depletion of DnaK in the Δ tig::kan mutant results in massive aggregation of cytosolic proteins. In Δ tig::kan cells, an increased amount of newly synthesized proteins associated transiently with DnaK. These findings show *in vivo* activity for a ribosome-associated chaperone, trigger factor, in general protein folding, and functional cooperation of this protein with a cytosolic Hsp70. Trigger factor and DnaK cooperate to promote proper folding of a variety of *E. coli* proteins, but neither

is essential for folding and viability at intermediate growth temperatures.

Because trigger factor can associate with ribosomes and nascent polypeptide chains, it is a prime candidate for a chaperone that is dedicated to assist folding of newly synthesized proteins. To elucidate its *in vivo* role genetically, we replaced the complete coding sequence of the *tig* gene (encoding trigger factor) with a kanamycin cassette (Fig. 1a). The Δ tig::kan mutant lacking trigger factor was viable and showed no growth defects between 15 and 42 °C in rich (Fig. 1b) and minimal media (not shown). Furthermore, Δ tig::kan mutants showed no defects in protein folding at 30 and 37 °C, as judged by unaltered specific activities of a reporter enzyme (firefly luciferase) and solubility of cellular proteins. Trigger factor is thus not essential for viability and protein folding in *E. coli*.

We investigated whether other chaperones compensate for the missing activity of trigger factor in Δ tig::kan cells by determining the phenotypes resulting from combining the Δ tig::kan allele with additional chaperone-gene mutations. The Δ tig::kan allele could be transduced with normal efficiency into the chromosomes of Δ hspG::lacZ (ref. 7), *secB::Tn5* (ref. 8), Δ ibpAB::kan (unpublished results) and temperature-sensitive *groEL* mutants (L140, L673 and L44)⁹ without apparent alteration of the existing growth phenotypes. However, it could not be introduced into Δ dnaK52 mutant cells¹⁰, which lack DnaK and have low levels of the DnaJ co-chaperone. This finding was further substantiated by co-transduction experiments in which a *Tn10* selective marker (*zba-3054::Tn10*) placed close to the Δ tig::kan allele was transduced into Δ dnaK52 and *dnaK*⁺ cells. The co-transduction frequency of the Δ tig::kan allele was 83% in *dnaK*⁺ cells and 0% in Δ dnaK52 mutant cells. Combination of the Δ tig::kan and Δ dnaK52 mutations thus causes synthetic lethality.

To investigate the cause of the synthetic lethality, we constructed Δ tig::kan and *tig*⁺ strains in which expression of the chromosomal *dnaK dnaJ* operon is under the control of the IPTG-regulatable promoter *P*_{A1/lacO-1} (*P*_{IPTG}*dnaKJ*, Fig. 1a)¹¹. On plates containing 1 mM IPTG, the *P*_{IPTG}*dnaKJ* Δ tig::kan and *P*_{IPTG}*dnaKJ* *tig*⁺ strains formed colonies at all temperatures tested (15, 30, 37 and 42 °C). On plates lacking IPTG, the *P*_{IPTG}*dnaKJ* *tig*⁺ cells did not form colonies at 15 and 42 °C (Fig. 1b), consistent with the cold-sensitive and heat-sensitive phenotype of Δ dnaK52 mutants¹². In the absence of IPTG, *P*_{IPTG}*dnaKJ* Δ tig::kan cells did not form colonies at any temperature tested (Fig. 1b); growth of these mutants at 30 and 37 °C was restored by expression of the *tig* gene from plasmids (not shown). The onset of synthetic lethality could also be monitored in liquid medium (Fig. 1c). After overnight growth in medium with IPTG, *P*_{IPTG}*dnaKJ* Δ tig::kan and *P*_{IPTG}*dnaKJ* *tig*⁺ cells were diluted into medium lacking IPTG. DnaK and DnaJ levels decreased from about twice wild-type levels to undetectable amounts after 8 h growth without IPTG (time point 3, Fig. 1c). *P*_{IPTG}*dnaKJ* Δ tig::kan cells depleted for DnaK and DnaJ started to show slower growth after 8 h (time point 3). At 8 and 10 h (time points 3 and 4), viability (determined as plating efficiency) decreased by 3 and 5 orders of magnitude, respectively, below the *P*_{IPTG}*dnaKJ* Δ tig::kan control grown with IPTG.

To account for synthetic lethality, we first ruled out that it resulted from DnaK's regulatory role as a negative modulator of σ ³², the transcriptional activator of chaperone and protease genes of the heat-shock regulon¹³. Accordingly, overproduction of chaperones and proteases resulting from DnaK and DnaJ depletion may be poisonous for Δ tig::kan cells. We artificially induced the heat-shock response in Δ tig::kan cells using a plasmid expressing the *rpoH* gene (encoding σ ³²) under the control of an IPTG-inducible promoter¹⁴. IPTG-induced chaperone overproduction was at least as high as chaperone overproduction caused by DnaK and DnaJ depletion (monitored, for example, for GroEL, Fig. 1d) but did not affect the viability of Δ tig::kan cells at 30 and 37 °C (Fig. 1d, lower panel).

To investigate whether synthetic lethality is caused by protein-

folding defects, we first tested the *in vivo* folding efficiency of firefly luciferase. At 5 h of growth without IPTG (time point 2, Fig. 1c) at 30 °C, DnaK and DnaJ levels were about 6-times lower than wild-type levels. Growth, viability and protein synthesis were not yet affected (data not shown). At this time point, synthesis of plasmid-encoded luciferase was induced for 30 min. Luciferase activity was reduced by 60–70% in $P_{IPTG}dnaKJ \Delta tig::kan$ cells as compared to $P_{IPTG}dnaKJ tig^+$ cells, although the amounts of luciferase present in both strains were similar (Fig. 2a). This decrease in specific activity was accompanied by increased luciferase aggregation (Fig. 2b).

We then tested the folding efficiency of *E. coli* bulk proteins. After 5 h of growth without IPTG (time point 2), cells were pulse-labelled with ^{35}S -methionine for 15 s and analysed for protein aggregation. Aggregated proteins and some membrane proteins were recovered by centrifugation and separated by two-dimensional-gel electrophoresis, followed by staining of pre-existing proteins and autoradiographic detection of newly synthesized proteins (Fig. 3). There was a large increase in the number (~40 proteins) and total amount (2.5-fold) of pelleted pre-existing proteins in $P_{IPTG}dnaKJ \Delta tig::kan$ cells as compared to $P_{IPTG}dnaKJ tig^+$ cells (Fig. 3a) and cells of both strains grown in IPTG (not

shown). Most spots that increased in intensity correspond to cytosolic proteins and thus represent aggregated proteins (for example, MetE and RpoB showed 4- and 25-fold increased aggregation, respectively; Fig. 3a, right panel). The amounts of membrane proteins present in the pellet fraction remained similar in the two strains tested (for example, inner and outer membrane proteins AtpD and OmpA; Fig. 3a), indicating that protein-folding defects affected predominantly non-membrane proteins. The autoradiogram revealed quantitatively (2.5-fold overall increase) and qualitatively similar changes in the solubility of newly synthesized proteins (Fig. 3b). Exposure of the gel was extended to allow detection of minor spots, but this also increased background staining, including that by nascent polypeptide chains.

The extent to which proteins misfold in DnaK- and DnaJ-depleted $\Delta tig::kan$ cells may be larger than detected, because proteases may degrade misfolded proteins and other chaperones may partially prevent aggregation. Furthermore, folding problems probably become more severe at later time points after IPTG withdrawal, conditions that were difficult to investigate because of general defects in, for example, protein synthesis.

To determine the functional relationship between DnaK and

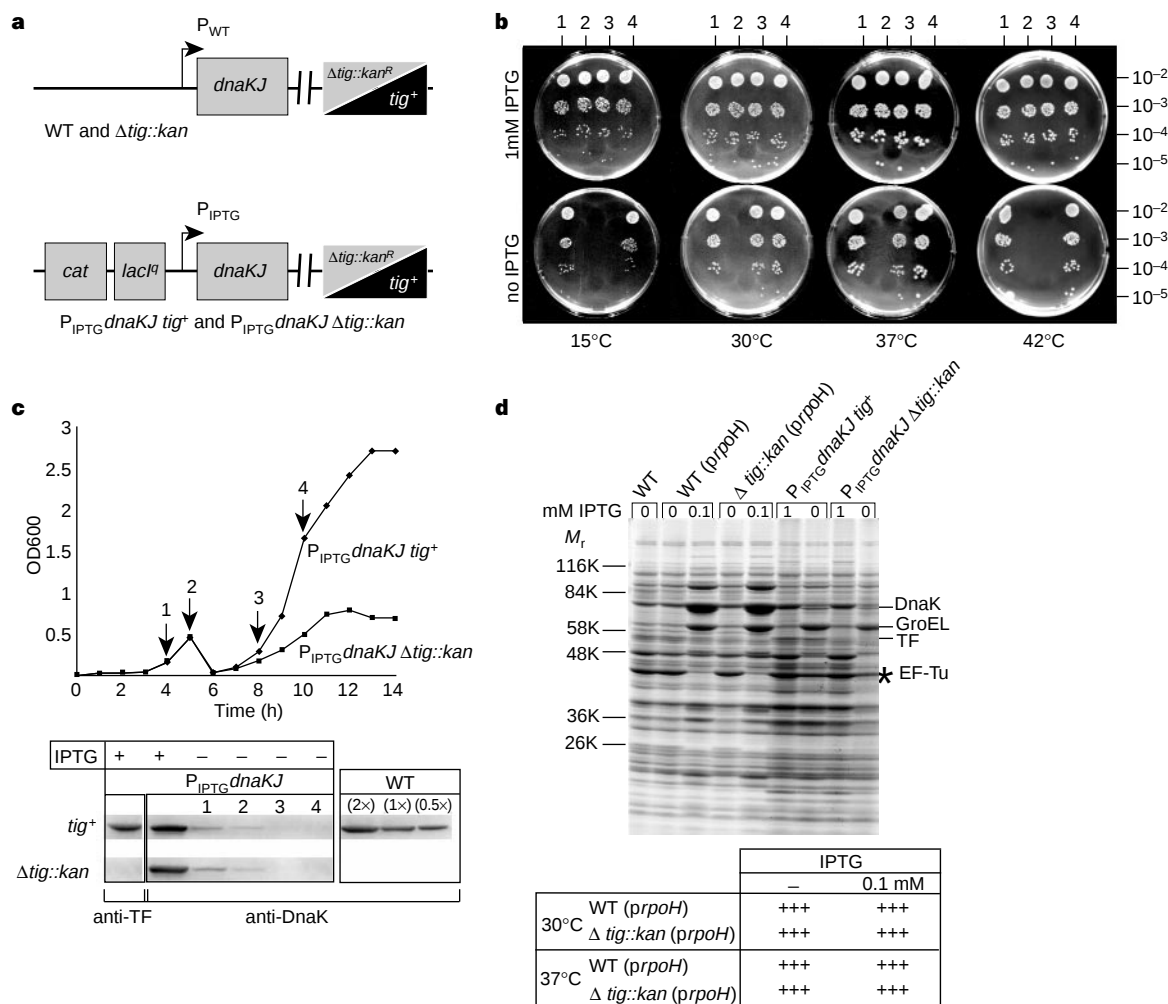


Figure 1 Synthetic lethality. **a**, Relevant genotypes of used strains. WT, wild type. **b**, Growth of wild-type (1), $P_{IPTG}dnaKJ \Delta tig::kan$ (2), $P_{IPTG}dnaKJ tig^+$ (3) and $\Delta tig::kan$ (4) cells on LB plates with or without IPTG at indicated temperatures for 24 or 96 h (15 °C). 10 μ l of diluted cultures were spotted. **c**, Growth of $P_{IPTG}dnaKJ tig^+$ and $P_{IPTG}dnaKJ \Delta tig::kan$ cells in LB liquid medium without IPTG at 37 °C. Top: at an OD₆₀₀ of 0.6, cells were diluted to maintain logarithmic growth. Arrows indicate time points of sample withdrawals. Similar growth behaviour was found for cells grown in M9 medium at 37 °C. Bottom: to detect DnaK and trigger factor levels during depletion, identical OD₆₀₀ equivalents were harvested at time points 1 to 4

(upper panel) and immunoblot analysis was performed using specific antisera. As controls, different OD₆₀₀ equivalents of wild-type cells were applied. **d**, Protein patterns of σ^{32} and *dnaK*, *dnaJ*-regulatable *tig*⁺ and $\Delta tig::kan$ cells. Top: Coomassie-stained SDS-PAGE of total lysates (identical OD₆₀₀ equivalents) of cultures grown in LB at 37 °C with IPTG or depleted from IPTG for 5 h. EF-Tu, for unknown reasons, is reduced in σ^{32} -overproducing cells. Bottom: growth of wild-type and $\Delta tig::kan$ cells with and without IPTG-induced σ^{32} overproduction on LB agar.

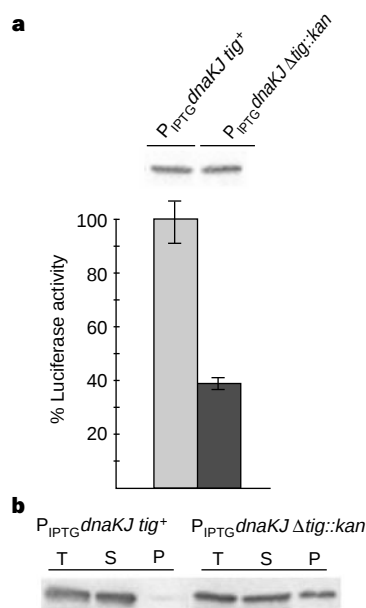


Figure 2 Specific activity and aggregation of firefly luciferase. **a**, *In vivo* activity and cellular levels of firefly luciferase in $P_{IPTG}dnaKJ\ tig^+$ (light-grey bar) and $P_{IPTG}dnaKJ\ \Delta tig::kan$ cells (dark-grey bar). DnaK- and DnaJ-depleted cells were induced for 30 min to produce luciferase and assayed for luciferase activity² (set 100% for $P_{IPTG}dnaKJ\ tig^+$ cells). Upper panel: identical OD₆₀₀ culture equivalents of the cells were subjected to immunoblot analysis with luciferase-specific antiserum. **b**, Solubility of luciferase. Lysates of cells treated as for **a** were prepared, divided into soluble and insoluble fractions by centrifugation and analysed by immunoblotting with luciferase-specific antiserum. Total lysate (T), supernatant (S) and aggregated material recovered in the pellet (P; double amount loaded).

trigger factor, we tested whether, in the absence of trigger factor, DnaK shows altered association with newly synthesized proteins tig^+ and $\Delta tig::kan$ cells were labelled for 15 s with ³⁵S-methionine and DnaK-bound substrates were rapidly recovered by co-immunoprecipitation with DnaK-specific antisera (Fig. 4a). About 2–4% of total labelled protein co-immunoprecipitated with DnaK from tig^+ cells, whereas 2–3-fold more labelled material was initially co-immunoprecipitated from $\Delta tig::kan$ cells (Fig. 4a, lower panel). The labelled material consisted of various proteins appearing as a smear (Fig. 4a), probably because nascent polypeptides were among these substrates. Chasing with non-radioactive methionine for 1–10 min before co-immunoprecipitation converted the smear to distinct bands, some of which associated with DnaK to similar extents in tig^+ and Δtig cells (Fig. 4a). Some nascent chains showed a rapid flux through DnaK (<1 min half-life), the extent, or perhaps the kinetics, of their association with DnaK being affected by trigger factor. Other proteins showed prolonged interaction with DnaK (>10 min half life), to an extent that is apparently not affected by trigger factor. The mechanistic basis for these protein classes remains unclear. Taking into account the efficiency of co-immunoprecipitation and the amount of recovered labelled protein, we estimate that 9–18% and 26–39% of newly synthesized proteins interact transiently with DnaK in wild-type and $\Delta tig::kan$ cells, respectively.

We also found that, in $\Delta tig::kan$ cells, a slightly increased amount (1.3-fold) of newly synthesized protein associates with GroEL (Fig. 4b). The most notable difference between the DnaK and GroEL substrates is an upper size limitation for GroEL substrates (relative molecular mass (M_r) of ~65K). This size exclusion by GroEL of large proteins, which are particularly prone to aggregation in $\Delta tig::kan$ mutants depleted for DnaK and DnaJ, explains the importance of the DnaK chaperone system for protein folding in $\Delta tig::kan$ cells.

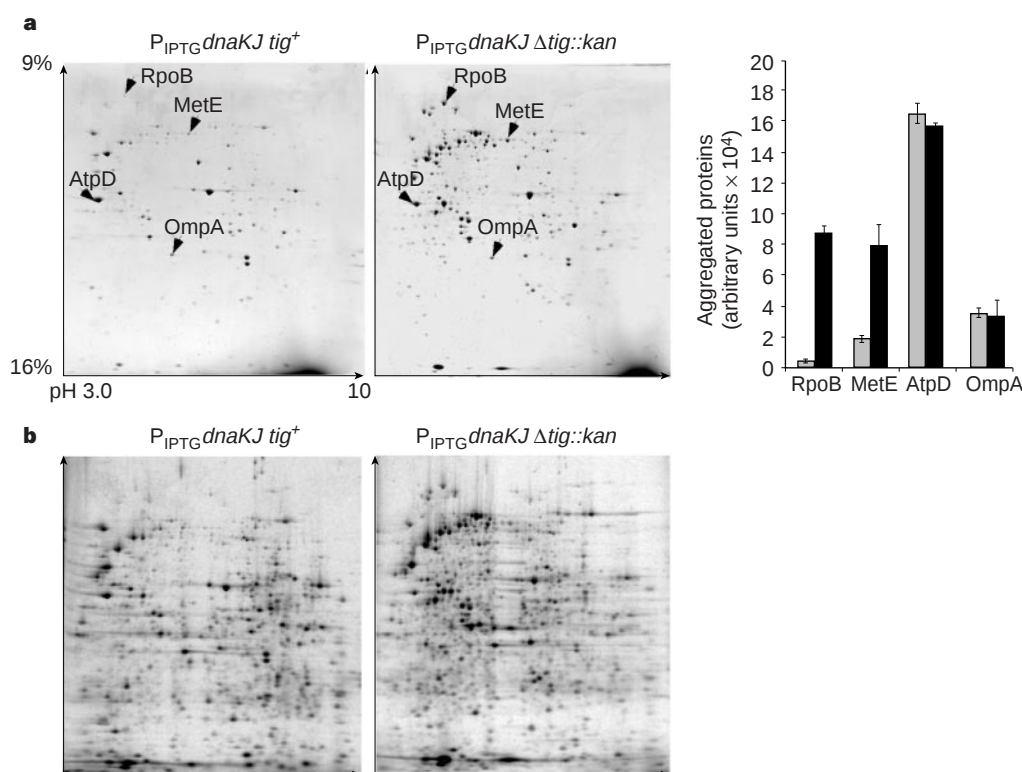


Figure 3 Aggregation of pre-existing and newly synthesized *E. coli* proteins. **a**, Coomassie-blue-stained gels of the pellet fractions of $P_{IPTG}dnaKJ\ tig^+$ and $P_{IPTG}dnaKJ\ \Delta tig::kan$ cells grown for 5 h in the absence of IPTG. Arrows indicate cytosolic (MetE, RpoB) and membrane (AtpD, OmpA) proteins identified by mass

spectrometry. Right: spot quantification of the Coomassie-stained gels. $P_{IPTG}dnaKJ\ tig^+$ cells, light-grey bars; $P_{IPTG}dnaKJ\ \Delta tig::kan$ cells, dark-grey bars. Quantification was performed using ImageMaster 2D program (Amersham/Pharmacia). **b**, Autoradiography of the Coomassie-stained gels shown in **a**.

Our findings indicate cooperative functions for trigger factor and DnaK in folding of newly synthesized proteins, although neither function is essential for folding. Trigger factor probably acts co-translationally, as it associates with ribosomes and nascent polypeptide chains^{3,15}. DnaK associates co- and/or post-translationally with a subset of newly synthesized proteins even in the presence of trigger factor, indicating that it regularly participates in protein folding even in wild-type cells. The nature of the functional cooperation between trigger factor and DnaK remains unclear. □

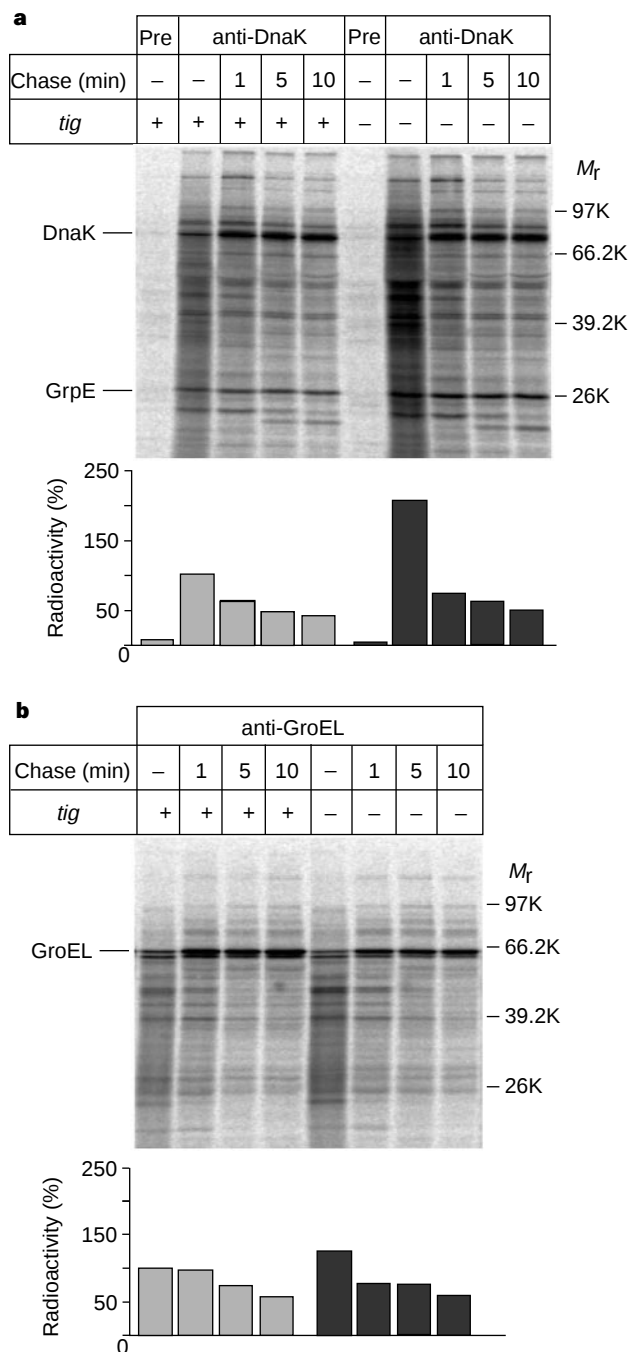


Figure 4 Association of DnaK with nascent polypeptide chains. **a**, Co-immunoprecipitation with DnaK-specific antisera. The lower panel shows quantification of radioactivity of the immunoprecipitated material above; *tig*⁺, light-grey bars; *Δtig::kan*, dark-grey bars. Radioactivity immunoprecipitated in the *tig*⁺ lysate after 15 s pulse is set 100%. **b**, Co-immunoprecipitation with GroEL-specific antibodies. The lower panel shows quantification of radioactivity of the immunoprecipitated material. Radioactivity immunoprecipitated in the *tig*⁺ lysate after 15 s pulse is set 100%.

Methods

Growth of P_{IPTG}*dnaKJ* cells. The strains used for depletion experiments are in the C600 background. Cells were grown overnight with 1 mM IPTG (isopropyl-β-D-thiogalactoside), then washed three times in saline to remove IPTG, inoculated in fresh medium to an absorbance at 600 nm (OD₆₀₀) of 0.015–0.03 and grown for analyses without IPTG at 37 °C for ~5 h to OD₆₀₀ = 0.6–0.8.

Luciferase activity. Cultures of P_{IPTG}*dnaKJ* *tig*⁺ and *Δtig::kan* cells containing plasmid pBB516 (ref. 11) were grown overnight at 30 °C in LB medium with ampicillin (100 μg ml^{–1}) containing 1 mM IPTG, washed three times in saline (0.8% NaCl) and inoculated in LB medium lacking IPTG to an OD₆₀₀ of 0.030. At mid-log phase, luciferase production was induced by the addition of 0.4% arabinose and *in vivo* activity was determined as described².

Isolation of insoluble proteins. Cultures were grown at 37 °C in M9 medium containing 0.2% glucose and all L-amino acids except methionine to an OD₆₀₀ of 0.6. Cells were pulsed with ³⁵S-methionine (15 μCi ml^{–1} final concentration) for 15 s, rapidly chilled on ice, lysed as described² and the protein content was determined (Bradford assay, Biorad). We subjected 2.4 mg protein of each lysate to a two-step centrifugation protocol where the first low-speed centrifugation removes most membrane proteins². 2D-gel electrophoresis of pellet fractions was performed according to the distributor (Amersham/Pharmacia). Experiments were reproduced at least three times.

Co-immunoprecipitation. Ten-millilitre aliquots of cells grown in M9 medium at 37 °C to mid-log phase were pulsed for 15 s with ³⁵S-methionine (15 μCi ml^{–1}; Amersham) and chased by addition of unlabelled methionine (160 μg ml^{–1}) when indicated. Cells were immediately transferred to 10 ml of frozen crushed solution (0.8% NaCl, 10 mM EDTA; pH 8.0), incubated for 5 min and then pelleted by low-speed centrifugation. Lysates were prepared as described², except that lysis buffer contained 100 mM EDTA to prevent ATP-dependent dissociation of DnaK-bound substrates, and that lysates were subjected to additional centrifugation for 5 min at 13,000g to remove insoluble material before immunoprecipitation. Eighty microlitres of lysate (2 mg ml^{–1}) was incubated with 80 μl protein A–Sepharose beads in PBS, 4 μl of appropriate antibodies and 1 ml of RIPA-buffer (50 mM Tris/HCl, 10 mM EDTA, 150 mM NaCl, 0.5% NP-40, pH 7.5) for 45 min at 8 °C. Beads were washed 3 times with RIPA-buffer, once with PBS/5 mM EDTA and finally boiled in 80 μl SDS-sample buffer. Incorporation of ³⁵S-methionine in TCA-precipitated lysates was determined to correct for differences in the labelling of the lysates.

Received 6 May; accepted 29 June 1999.

1. Ewalt, K. L., Hendrick, J. P., Houry, W. A. & Hartl, F. U. *In vivo* observation of polypeptide flux through the bacterial chaperonin system. *Cell* **90**, 491–500 (1997).
2. Hestekamp, T. & Bukau, B. Role of the DnaK and HscA homologs of Hsp70 chaperones in protein folding in *E. coli*. *EMBO J.* **17**, 4818–4828 (1998).
3. Hestekamp, T., Hauser, S., Lütcke, H. & Bukau, B. *Escherichia coli* trigger factor is a prolyl isomerase that associates with nascent polypeptide chains. *Proc. Natl Acad. Sci. USA* **93**, 4437–4441 (1996).
4. Valent, Q. A. *et al.* Early events in preprotein recognition in *E. coli*: interaction of SRP and trigger factor with nascent polypeptides. *EMBO J.* **14**, 5494–5505 (1995).
5. Scholz, C., Stoller, G., Zarnt, T., Fischer, G. & Schmid, F. X. Cooperation of enzymatic and chaperone functions of trigger factor in the catalysis of protein folding. *EMBO J.* **16**, 54–58 (1997).
6. Stoller, G. *et al.* A ribosome-associated peptidyl-prolyl *cis/trans* isomerase identified as the trigger factor. *EMBO J.* **14**, 4939–4948 (1995).
7. Bardwell, J. C. A. & Craig, E. A. Ancient heat shock gene is dispensable. *J. Bacteriol.* **170**, 2977–2983 (1988).
8. Kumamoto, C. A. & Beckwith, J. Mutations in a new gene, *secB*, cause defective protein localization in *Escherichia coli*. *J. Bacteriol.* **154**, 253–260 (1983).
9. Zeilstra-Ryalls, J., Fayet, O., Baird, L. & Georgopoulos, C. Sequence analysis and phenotypic characterization of *groEL* mutations that block λ and T4 bacteriophage growth. *J. Bacteriol.* **175**, 1134–1143 (1993).
10. Paek, K. H. & Walker, G. C. *Escherichia coli* *dnaK* null mutants are inviable at high temperature. *J. Bact.* **169**, 283–290 (1987).
11. Tomoyasu, T., Ogura, T., Tatsuta, T. & Bukau, B. Levels of DnaK and DnaJ provide tight control of heat shock gene expression and protein repair in *E. coli*. *Mol. Microbiol.* **30**, 567–581 (1998).
12. Bukau, B. & Walker, G. C. Cellular defects caused by deletion of the *Escherichia coli* *dnaK* gene indicates roles for heat shock protein in normal metabolism. *J. Bacteriol.* **171**, 2337–2346 (1989).
13. Gross, C. A. in *Escherichia coli and Salmonella* (ed. Neidhardt, F. C.) 1382–1399 (ASM, Washington, 1996).
14. Gamer, J. *et al.* A cycle of binding and release of the DnaK, DnaJ and GrpE chaperones regulates activity of the *E. coli* heat shock transcription factor σ³². *EMBO J.* **15**, 607–617 (1996).
15. Valent, Q. A. *et al.* Nascent membrane and presecretory proteins synthesized in *Escherichia coli* associate with signal recognition particle and trigger factor. *Mol. Microbiol.* **25**, 53–64 (1997).

Acknowledgements. We thank members of the Bukau lab, T. Hestekamp, M. Müller and H.-G. Koch for comments on the manuscript and discussions. This work was supported by grants of the DFG and the Fonds der Chemie to B.B.

Correspondence and requests for materials should be addressed to B.B. (e-mail: bukau@sun2.ruf.uni-freiburg.de).